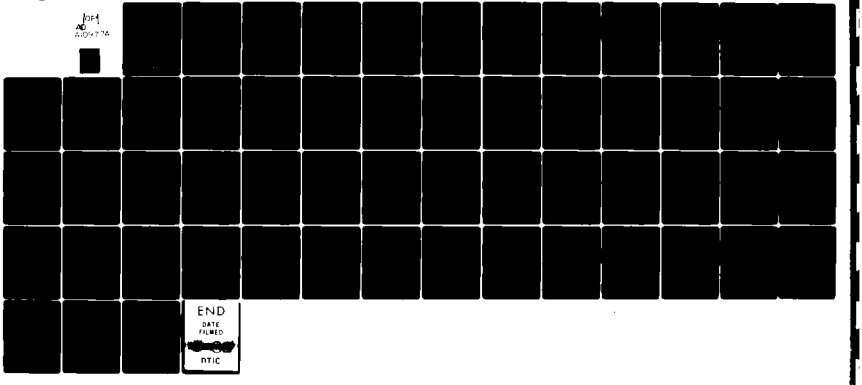
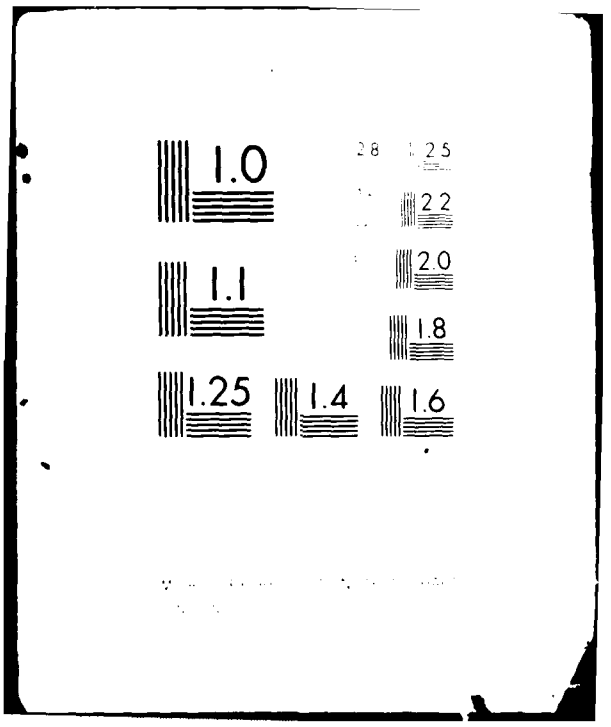
AD-A109 774 STANFORD UNIV CA DEPT OF OPERATIONS RESEARCH F/G 12/1
THE SHIFT-FUNCTION APPROACH FOR MARKOV DECISION PROCESSES WITH --ETC(U)
JUL 81 S STIDHAM, J VAN NUNEN N00014-76-C-0418
UNCLASSIFIED TR-98 NL




1.0
1.1
1.25
1.4
1.6
2.0
1.8
2.2

LEVEL II

14

THE SHIFT-FUNCTION APPROACH FOR MARKOV DECISION PROCESSES WITH UNBOUNDED RETURNS

BY

SHALER STIDHAM, JR. and JO VAN NUNEN

TECHNICAL REPORT NO. 98
JULY 1981

PREPARED UNDER CONTRACT
N00014-76-C-0418 (NR-047-061)
FOR THE OFFICE OF NAVAL RESEARCH

Frederick S. Hillier, Project Director

DEPARTMENT OF OPERATIONS RESEARCH
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

THE SHIFT-FUNCTION APPROACH FOR MARKOV DECISION PROCESSES WITH UNBOUNDED RETURNS

BY

Shaler Stidham,Jr.* and Jo van Nunen**

TECHNICAL REPORT NO. 98
JULY 1981

PREPARED UNDER CONTRACT
N00014-76-C-0418 (NR-047-061)
FOR THE OFFICE OF NAVAL RESEARCH
Frederick S. Hillier, Project Director

DEPARTMENT OF OPERATIONS RESEARCH
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

This research was supported in part by National Science Foundation Grant ECS 80-17867 Department of Operations Research, Stanford University and issued as Technical Report No. 60.

*The research of this author was partially supported by National Science Foundation Grant No. ENG 78-24420 at North Carolina State University.

**Graduate School of Management, Delft, The Netherlands. The research of this author was done during an appointment as Visiting assistant Professor of Operations Research and Mathematics, North Carolina State University, January to June, 1978.

The first draft of this report was titled "Uniform Convergence of Successive Approximations in Dynamic Programming with Unbounded Returns and Non-Zero Terminal Value Function".

0. Introduction

We consider a Markov decision process with a general state space and general action space. The system is observed at discrete points in time. If it is in state s and action a is taken, then an immediate return r(s,a) is earned and the system makes a transition to a new state according to the (possibly defective) transition probability measure p(s,a;·). The objective is to maximize the expected total return over a finite or infinite horizon from each possible starting state. Discounting is accommodated by incorporating the discount factor in the transition probabilities.

We seek a set of realistic and easily verified conditions under which optimal policies can be computed (or approximated) in an efficient way. What distinguishes our approach from many others in the literature is that the conditions we develop are appropriate to a specific class of Markov decision processes: those that arise in the control of stochastic service and storage systems, such as queueing, replacement, and inventory systems. Under these conditions we are able to establish the standard results of the theory of Markov decision processes: (i) that the optimal value function satisfies the optimality equation of dynamic programming, (ii) that it is the unique solution in a certain class of functions, (iii) that a stationary policy attaining the maximum in the optimality equation is optimal among all policies, and (iv) that the method of successive approximations converges (in other words, the finite-horizon optimal value functions approach the infinite-horizon optimal value function).

It is customary in the literature to start with a general Markov decision model and impose regularity conditions, such as e.g. uniform, polynomial, or exponential bounds on the return functions [3], [14], [20], [37], as they are needed to derive desired results. By contrast, we begin with what we feel to be appropriate abstractions of the applications in which we are interested. We

impose on our model a set of conditions, mostly involving monotonicity of return functions and transition probabilities, that are common to many of the control models for queueing, replacement, and inventory systems in the literature, and then work toward achieving as many of the goals (i) - (iv) as possible. In fact, we are able to achieve all four goals. Moreover, we are able to do this, by means of certain transformations, in the context of an approach based on contraction mappings with respect to the sup norm, so that the convergence of successive approximations is uniform and geometric. This property makes it possible to use various techniques to accelerate convergence, such as bounds, elimination of sub-optimal actions, and transformations to reduce the spectral radius of the process (see, for example, [8], [9], [17], [18], [19], [20], [22]).

In Section I we introduce our basic Markov decision model and establish notation. The special cases of this model studied in this paper are all basically examples of the essentially negative model of Hinderer [11]. That is, the one-stage return function is bounded above and the one-stage discount factor is strictly smaller than one. Like Schäl [24] we allow the discount factor to depend on the state and action, so that our model covers semi-Markov decision processes. Our notation incorporates the discount factor in the transition probabilities and allows for defective transition probabilities (cf. [20], [21]). For n-stage problems we extend the Schäl model to allow a non-zero terminal-value (scrap) function (cf. [10], [32]). In the context of the successive-approximations method for infinite-stage problems, this is equivalent to allowing a non-zero starting function.

In Sections 2 and 3 we study two special cases of the basic decision model of Section 1, both of which abstract some of the properties commonly found in stochastic service and storage systems. The model of Section 2 allows unbounded return functions, but the structure of the return function and transition

probabilities gives rise to bounded optimal value functions. Hence goals (i) - (iv) can be achieved and, moreover, the convergence of the method of successive approximations is uniform with respect to the sup norm. Although the conditions of this model may seem restrictive, we give an example from queueing control to show that they can be satisfied in some applications.

The model of Section 3 also allows unbounded return functions, but places fewer restrictions on the return functions and transition probabilities and hence is applicable to a wider class of Markov decision processes. In fact the conditions of this model seem to be satisfied by nearly all the queueing-control models considered in the literature. We give some examples in support of this assertion and also present an inventory model in which our conditions are satisfied. The assumptions for this model are weaker than those of the models in the literature in which an (s,S) policy is shown to be optimal. As in the case of the model of Section 2, we are able to achieve goals (i) - (iv) for the model of this section, and establish uniform convergence of successive approximations. The optimal value functions are not bounded for this model, however. By means of a shift transformation [7], [20], [21] we show how to convert this model to an equivalent model satisfying the conditions of Section 2 and give an economic interpretation of the transformation. As a shift function we propose (among other possibilities) the infinite-horizon value function from a particular reference policy of a simple form. In queueing-control models the reference policy is usually an extremal policy, e.g., the policy that rejects all customers in an arrival-control problem or the policy that always uses the maximal service rate in a service-rate-control problem. In the inventory-control model the reference policy could be the policy that orders nothing when inventory is positive and orders up to zero, if possible, when the inventory is negative.

1. Basic Decision Model

We now give a detailed description of the basic Markov decision model and establish the notation that will be used throughout the paper. Informally speaking, the object of study is a system that can be controlled at discrete time points, called stages and labeled $t = 0,1,2,\ldots$. If at stage t the system is observed to be in state $s \in S$, the decision maker can select an action $a \in D(s)$, the set of admissible actions. If he selects action a in state s at stage t, then an immediate return $r(s,a)$ is earned and the state at stage $t + 1$ will be in the set B with probability $p(s,a; B)$.

As in [20],[21] we allow for defective transition probabilities, i.e. $p(s,a;S) < 1$. The model thus includes discounted Markov and semi-Markov decision processes, as well as stopping problems. In a more formal development, the model can be converted to one with proper transition probabilities by introducing an absorbing state (cf. [20], [21]). As this conversion is by now standard in the literature, we shall assume that it has been done and make no further reference to it.

A policy π is defined in the usual way [3], [4], [11], [24], [26] as a collection of decision rules for choosing actions at each stage t. A policy is called stationary, and denoted simply by f, if it always chooses the same action, $a = f(s) \in D(s)$, whenever the system is in state $s \in S$. The set of all stationary policies will be denoted by F. Each starting state $s \in S$ and policy π, together with the transition probability measure p, determine a stochastic process $\{(X_t, A_t), \quad t = 0,1,\ldots\}$ with associated probability measure P_s^π . Here X_t denotes the state and A_t the action at stage t. We shall denote by $E_s^\pi [\cdot]$ the expectation operator associated with P_s^π , and write $E^\pi [\cdot]$ to denote the

function that assigns value $E_s^\pi[\cdot]$ to the point $s \in S$.

In order to keep the exposition simple, we shall make no reference in this paper to measure-theoretic and topological conditions needed to ensure that the stochastic processes and associated measures P_s^π are well defined. Our development is rigorous if, for example, all the sets referred to above are standard Borel spaces and the functions are measurable. (See, e.g., Hinderer [11], Schäl [24], Serfozo [26], or Stidham [32] for details.)

Associated with each policy π, formally define the infinite -horizon value function V^π by

$$V^\pi(s) := E_s^\pi [\sum_{t=0}^{\infty} r(X_t, A_t)] \qquad (s \in S)$$

and define the infinite - horizon optimal value function V* by

$$V^*(s) := \sup_\pi V^\pi(s) \qquad (s \in S)$$

In order to ensure that the expectations are well defined, we shall make specific assumptions regarding r and p for each of the specific models considered in this paper. In all cases the models will be special cases of an essentially negative model (cf Hinderer [11]).

For finite-stage problems we allow a terminal-value (scrap) function, $V_o: S \to \underline{R}$. That is, if the horizon length is n, then the system terminates upon reaching stage t=n and earns a terminal value $V_o(X_n)$. Associated with each policy π, formally define the n-stage value function V_n^π by

$$V_n^\pi(s) := E_s^\pi[\sum_{t=0}^{n-1} r(X_t, A_t) + V_o(X_n)] \qquad (s \in S)$$

and the n-stage optimal value function V_n^* by

$$V_n(s) := \sup_{\pi} V_n^{\pi}(s) \qquad (s \in S)$$

Again, we shall impose specific conditions on r, p, and V_o to ensure that these functions are well defined for each of the models considered.

For a stationary policy f define the operator P(f) by

$$(P(f)v)(s) := \int p(s,f(s); ds')v(s') \qquad (s \in S)$$

for all functions v such that the integral is well defined. Similarly, define the operator L(f) by

$$L(f)v := r(f) + P(f)v,$$

where r(f) is the function whose value for the argument s is r(s,f(s)), $s \in S$. It follows from the definitions that, for a stationary policy f,

$$V^f = L(f)V^f = r(f) + P(f)V^f, \qquad (1.1)$$

$$(L(f))^n 0 \to V^f, \text{ as } n \to \infty. \qquad (1.2)$$

Define the operator U by

$$Uv := \sup_{f \in R} . L(f)v.$$

Under the conditions satisfied by the models considered in this paper, it can be shown [24], [32] that V* and V_n satisfy the optimality equations

$$V^* = UV^* \qquad (1.3)$$

$$V_n = UV_{n-1} = U^n V_o, \qquad n > 1 \tag{1.4}$$

As in Stidham [32] our main goal in this paper will be to establish conditions (on r, p, and V_o) under which successive approximations converges, that is, $V_n \to V^*$, In contrast to [32], in which pointwise convergence was established using extensions of the methods of Strauch [34], Hinderer [11], and Schäl [24], we shall here seek uniform convergence (convergence in sup norm) under conditions in which the operator P(f) is contractive, but r(f) is unbounded. As indicated in the introduction, our conditions are specifically tailered to fit control models for stochastic service and storage systems.

For any function v: $S \to R$, define the supremum norm of v, denoted $\| v \|$, by

$$\| v \| := \sup_{s \in S} |v(s)|$$

Let $\underline{R} := R \cup \{-\infty\}$ and let V: $S \to R$ be a given function (hereafter called a reference function). Let $W(V)$ be the Banach space of all functions uniformly bounded away from V. That is,

$$W(V) := \{v: \ S \to \underline{R} \ | \ \| v-V \| < \infty\}.$$

The following lemma, which follows from (1.3) and the contraction-mapping fixed-point theorem, will be used frequently in our analysis.

Lemma 1.1. Suppose

(i) $V^* \in W(V)$

(ii) U is contractive on $W(V)$, i.e., there exists a ρ, $0 \leq \rho < 1$, such that

$$\| Uu - Uv \| \leq \rho \| u - v \| \quad \text{for all } u, v \in W(V).$$

Then, for all $v \in W(V)$, as $n \to \infty$,

$$\| U^n v - V^* \| \leq \rho^n \| v - V^* \| \to 0$$

and V^* is the unique fixed point of U in $W(V)$.

(It should be noted that (i) and (ii) imply that U: $W(V) \to W(V)$.)

2. Model I

In this section we study a special case of the basic decision model presented in the previous section. The special structure of this model makes it possible to apply the theory of contraction mappings to deduce uniform geometric convergence of the method of successive approximations, even though the one-period return function may be unbounded. In fact, our simple assumptions describe an important class of Markov decision processes for which the classical theory of Markov decision processes, which was developed under conditions that only allowed uniformly bounded one-stage returns, is applicable. The difference between our conditions and the classical conditions (cf., e.g., [3], [5]) is that we do not require r(f) to be uniformly bounded for all f, but only for certain f. Our conditions may in fact be part of the folklore, but we could not find them in the literature.

Although the assumptions of our model may at first seem artificial and restrictive, they are satisfied in certain applications, as we shall demonstrate in the latter part of this section by means of an example from the control of queues. Of more significance, perhaps, is the fact that for a wide class of decision models, including most of the queueing-control models in the literature, it is possible to transform the problem into an equivalent problem that satisfies the assumptions of this section. We shall demonstrate this transformation in the next section.

Let $e: S \to R$ be the unit function; that is, $e(s) \equiv 1$. For each stationary policy f, define the supremum norm of the transition operator P(f) by

$$\| P(f) \| := \sup_{s \in S} (P(f)e)(s)$$

$$= \sup_{s \in S} p(s, f(s), S).$$

Also define $\bar{r}: S \to \mathbb{R} \cup \{+\infty\}$ by

$$\bar{r} := \sup_{f \in F} r(f)$$

We shall need the following conditions.

Condition 2.1. $\rho := \sup_{f \in F} \| P(f) \| < 1$

Condition 2.2 $M := \| \bar{r} \| < \infty$

For discounted problems Condition 2.1 is equivalent to having the discount factor uniformly strictly less than one. Condition 2.2 implies (but is not implied by) having the one-stage return r(f) uniformly bounded above. Thus Model I is a special case of an essentially negative dynamic-programming model. Note that Condition 2.2 does not require the one-stage returns to be bounded below, so that Model I is not a special case of the classical discounted bounded-return model of Blackwell and Denardo. However, we do require r(f) to be uniformly bounded for the subclass of myopic stationary policies, that is, policies f for which $r(f) = \bar{r}$.

In this section we shall be interested in the Banach space $U = \mathcal{W}(0)$ of all uniformly bounded functions, that is,

$$\mathcal{W} := \{v: s \to R \mid \| v \| < \infty \}$$

The key results of this section are contained in the following theorem.

Theorem 2.1. Assume Conditions (2.1) and (2.2). Then $V^* \in \mathcal{W}$ is the unique bounded solution to the optimality equation

$v = Uv$

and, for any $V_o \in W$, $\| V^* - V_n \| \leq \rho^n \| V^* - V_o \| \leq \rho^n [(1 - \rho)^{-1} + \| V_o \|]$ so that $V_n = U^n V_o \to V^*$ uniformly and geometrically (with respect to the sup norm).

Proof. These results follow from Lemma 3.3 and Theorem 3.1 in van Nunen and Wessels [21], extended in an obvious way to a general (not necessarily countable) state space, since Conditions 2.1 and 2.2 imply assumption 2.3 of [21].

Alternatively, the theorem may be proved directly by verifying that (1) and (1) of Lemma 1.1 hold, so that the classical contraction-mapping theory applies.

Conditions 2.1 and 2.2 give a simple special case in which the very general conditions of van Nunen and Wessels [21] hold, and hence the theory of contraction mappings based on weighted sup norms can be applied. Indeed, from the fact (demonstrated in Theorem 2.1) that only "good" (i.e. myopic) policies need to have bounded return functions in order for the model to be contractive, it follows that the classical theory [3], [5] based on ordinary sup norms is applicable.

The advantage of our conditions over the more general conditions in [21] is that they are simple to state and can be easily checked in applications. They would not be of much use, however, if they were too restrictive to have any significant application (except trivially in the case where $r(\cdot,\cdot)$ itself is bounded.). In the remainder of this section, we shall illustrate the application of Conditions 2.1 and 2.2 to problems with special structure, which makes verification of these conditions easy. To this end we shall present new conditions that imply Conditions 2.1 and 2.2, in the context of a general model with partially ordered state and/or action spaces. These conditions, which involve monotonicity of r and p, are much stronger than necessary for conditions 2.1 and 2.2, and consequently lead to sharper results than Theorem 2.1. They are related to but weaker than conditions that have been proposed in the literature (cf, e.g., Stidham and Prabhu [33], Serfozo [26]) for a different purpose: namely, showing that an optimal policy has a particular (e.g., monotonic or control-limit)

form. We illustrate the application of these stronger conditions with an example from control of queues. In the next section we show how more general problems can often be transformed into equivalent problems having this structure, so that the results of this section can be applied.

For the remainder of this section, suppose that the state space S is partially ordered by a relation "$\leq$" and that $D(s) = A$, for all $s \in S$. A function $v: S \to R$ is called increasing (decreasing) if $v(s) \leq v(t)$ ($v(s) \geq v(t)$) for all $s \leq t$ in S. A set $B \subset S$ is called increasing (decreasing) if $s \in B$ implies $t \in B$ whenever $s \leq t$ ($s \geq t$). We shall need the following conditions.

Condition 2.3. There is an element $0 \in S$ that is minimal with respect to the relation "$\leq$". That is, $s \geq 0$ for all $s \in S$.

Condition 2.4. For all $a \in A$, $r(s,a)$ is decreasing in $s \in S$; moreover $\sup_{a \in A} r(0,a) =: M < \infty$.

Condition 2.5. For each $s \in S$, there exists an $a \in A$ such that $r(s,a) \geq 0$.

Conditions 2.3 and 2.4 imply that $r(s,a) \leq M < \infty$, for all $s \in S$, $a \in A$. Condition 2.5 implies that $\bar{r}(s) \geq 0$, for all $s \in S$. Hence $\| \bar{r} \| < M < \infty$ and Condition 2.2. applies. Moreover, $0 \leq V^* \leq M(1-\rho)^{-1} < \infty$. Hence we have the following corollary of Theorem 2.1.

Corollary 2.2. Assume Conditions 2.1 and 2.3, 2.4, 2.5. Then $0 \leq V^* \leq M(1-\rho)^{-1}) < \infty$. V^* is the unique solution in U to the optimality equation

$$v = Uv$$

and, for any $V_0 \in U$, $V_n = U^n V_0 \to V^*$ uniformly and geometrically with respect to the sup norm.

The monotonicity of r implied by Condition 2.4 makes it natural to look for conditions under which V_n and V^* will be monotonic functions on S. (Monotonicity of V_n and V^* is often used in proving that an optimal policy has a particular form,

for Markov decision processes with special structure. See, for example, Sobel [28], Stidham and Prabhu [31], and Serfozo [26]). For this purpose we shall need the following additional condition.

Condition 2.6. For all $s \in S$, $s' \in S$, $s \leq s'$ implies $p(s, a, B) \leq p(s', a, B)$ for $a \in A$ and each increasing set $B \subset S$.

When Condition 2.6 holds we say that p is stochastically increasing in s (cf. Lehman [13], Bessler and Veinott [2], Serfozo [26]. Serfozo calls such p a monotone transition probability). From Serfozo [26] we get the following.

Lemma 2.3. Condition 2.6 holds if and only if, for every stationary policy f such that $f(s) \equiv a$, for all $s \in S$, and every increasing (decreasing) function $v: S \to \underline{R}$, $P(f)v$ is increasing (decreasing).

Now the following theorem can be proved easily by induction on n.

Theorem 2.4. Assume Conditions 2.1, and 2.3 - 2.6. If $V_0 \in W$ and is decreasing, then $V_n \in W$ and is decreasing, for each $n \geq 1$, $V_n \to V^*$ uniformly and geometrically, and V^* is decreasing and the unique solution in W to $v = Uv$.

Remark 2.1. Monotone return functions and transition probabilities are encountered often in applications of Markov decision processes, particularly to queueing and replacement systems. Such systems are often closely related to random walks and thus have a (nearly) additive transition structure. Specifically, the state represents the "quantity" (e.g. number of customers) in the system and transitions occur by means of inputs and outputs, which are (nearly) independent of the current state and either or both of which may be subject to control (see Stidham and Prabhu [31], Section 4.1). Such a transition structure typically gives rise to monotone transition probabilities. (For a simple example in the context of a controlled random walk, see Serfozo [26], Section 5.) The basic

idea, of course, is that whatever the (fixed) action, the more quantity in the system now, the more is likely to be in the system at the next stage. The fact that the return function, $r(\cdot,a)$, is decreasing comes about because there is usually a cost (e.g., inventory holding cost, customer waiting cost) associated with having quantity in the system: the more the quantity, the higher the cost.

Remark 2.2. The observation that $V^* \geq 0$ (see Corollary 2.2) depended on the fact that Condition 2.5 implies the existence of a stationary policy that never takes an action leading to a negative immediate return. At first glance it might seem that a stronger statement could be made, namely, that without loss of optimality one may restrict attention to policies that never take an action a from any state s such that $r(s,a) < 0$. If this were true, then the problem would be equivalent to one in which $r(s,a) \geq 0$, for all $s \in S$, $a \in A$. It is easy to construct a counterexample, however, to show that Condition 2.5 does not imply that actions leading to negative immediate returns can be ignored. The problem, of course, is that it may be advantageous to incur a negative return now in order to get into a set of states with large positive returns.

There are, however, realistic additional conditions under which our model is equivalent to one with a non-negative return function. As an example we offer

Condition 2.7. The action space A is partially ordered by a relation "$\leq$". For all $s \in S$, $a \in A$ such that $r(s,a) < 0$, there exists an $a' \in A$, $a' < a$, such that $r(s,a') \geq 0$. For each $s \in S$, $p(s,a\ ;\ \cdot)$ is stochastically increasing in $a \in A$. (For an application in which this assumption holds, see below.)

Theorem 2.5. Assume Conditions 2.1, 2.3-2.7. Suppose $V_0 \in \mathcal{W}$ is decreasing. Then, for each $n \geq 1$, $V_n \in \mathcal{V}$, is decreasing, and satisfies the restricted optimality equation

$$V_n = \sup_{f \in F_+} L(f)\, V_{n-1}\ , \tag{2.1}$$

where $F_+ := \{f \in F \mid r(f) \geq 0\}$.

Moreover, $V_n \to V^*$ uniformly and geometrically, and V^* is decreasing and the unique solution in $\mathcal{V}$ to the restricted optimality equation

$$V^* = \sup_{f \in F_+} L(f) V^* \tag{2.2}$$

If $V_o \equiv 0$, then the convergence of V_n to V^* is monotonic : $V_n \geq V_{n-1}$, $n \geq 1$.

Proof. In light of Theorem 2.4, it suffices to show that the supremum over all stationary policies f in the original statement of the optimality equations can be replaced by a supremum over F_+ without loss of optimality. We have, for example,

$$V^* = UV^* = \sup_{f \in F} \{r(f) + P(f)V^*\}.$$

Let f be an arbitrary stationary policy. It follows from Condition 2.7 that there exists a policy $f' \in F_+$ with $f' \leq f$ and $r(f') \geq r(f)$. (Set $f'(s) = f(s)$ if $r(s,f(s)) \geq 0$; for $s \in S$ such that $f(s) = a$ and $r(s,a) < 0$, set $f'(s) = a'$, where $a' < a$ and $r(s,a') \geq 0 > r(s,a)$.) Since V^* is decreasing (by Theorem 2.4) it follows from the second part of Condition 2.7 that $P(f) V^* \leq P(f') V^*$. (The situation parallels that covered by Lemma 2.3, except that now we are dealing with a stochastic ordering on A rather than S.) Therefore,

$$r(f') + P(f') V^* \geq r(f) + P(f) V^*$$

and consequently, since f was arbitrary and $f' \in F_+$,

$$\sup_{f \in F_+} L(f) V^* \geq \sup_{f \in F} L(f) V^* \geq \sup_{f \in F_+} L(f) V^*,$$

so that (2.2) holds. The proof of (2.1) is formally identical. Monotonicity of convergence when $V_o \equiv 0$ follows by induction on n, using the fact that U is a monotone operator and $UV_o = \bar{r} \geq 0 = V_o$.

A queueing-control application

Consider the following model for control of arrivals to a generalized queueing system, which includes several arrival-control models in the literature as special cases (see Johansen and Stidham [12] for details). Customers arrive at intervals which are independent and identically distributed as a random variable T, with $\Pr\{T > \varepsilon\} > 0$ for some $\varepsilon > 0$. Each customer brings a certain amount of potential input to the system. The potential inputs of successive customers are independent and identically distributed as a random variable S. At each arrival instant the system controller has the option of accepting or rejecting the entire potential input of the arriving customer. If an input is accepted then it is added to the quantity, s, in the system and a net benefit, b(s), is earned. We assume that $b(s) = r - C(s)$, where r is a reward, or utility of service, and C(s) is a waiting cost, a non-decreasing function of s. If an input is rejected, the net benefit is 0. Potential output from the system is governed by an uncontrollable stochastic process with non-negative stationary independent increments, $\{N(\tau), \tau \geq 0\}$. Thus, at the end of a time interval of length τ, which begins with a quantity s in the system and during which no arrivals are accepted, the quantity in the system is distributed as $(s - N(\tau))^+$. Future benefits are continuously discounted at rate $\alpha > 0$.

In reference [12] the reward is allowed to be a random variable, but in all other respects the model is the same. The model specializes to a GI/G/1 system with quantity interpreted as work in the system, if S has the distribution of the service time and $N(\tau) \equiv \tau$ (see [33], [6]). If $S \equiv 1$ and $\{N(\tau), \tau \geq 0\}$ is a Poisson process, then the model specializes to a GI/M/1 system with quantity interpreted as the number of customers in the system (see [16], [31]).

The problem of maximizing the expected discounted total net benefit over a finite or infinite horizon can be formulated as a special case of our Markovian

decision model, in which the state s is the quantity found in the system by an arrival, the action a = 1 (0) denotes acceptance (rejection) of a customer, $r(s,a) = a(r-C(s))$, and $p(s,a;\ B) = E[e^{-\alpha T}\ \underset{\sim}{1}((s + a\ S - N(T))^{+} \in B)]$, where α is the continuous-time discount rate and $\underset{\sim}{1}(E)$ is the indicator of the event E. The finite-horizon and infinite-horizon optimality equations are, respectively,

$$V_n(s) = \max_{a=0,1} \{a(r-C(s)) + E[e^{-\alpha T}\ V_{n-1}((s + a\ S - N(T))^{+})]\},$$

$n \geq 1$ (assume $V_0 \equiv 0$), and

$$V^*(s) = \max_{a=0,1} \{a(r-C(s)) + E[e^{-\alpha T}\ V^*((s + a\ S - N(T))^{+})]\} \qquad (2.3)$$

$s \in S = [0, \infty)$. It is easily verified that Assumptions 2.1-2.7 are satisfied. Hence Theorem 2.5 applies, which implies in particular that it is optimal at each stage n, $1 \leq n \leq \infty$, to reject the arriving customer (a = 0) if $r < C(s)$, that is, if his individual net benefit is negative. The converse is not generally true (except for n = 1): an optimal policy may reject a customer even though $r \geq C(s)$, that is, even though it is in his individual interest to join. (For further discussion of this phenomenon, see Johansen and Stidham [12] and the references cited therein.) Theorem 2.5 also implies that V_n and V^* are non-negative and non-increasing and that the convergence of V_n to V^* is monotonic ($V_n \geq V_{n-1}$) as well as uniform and geometric.

3. Model II

In this section we consider another special case of the general decision model of Section 1. Unlike the model of the previous section, this model does not have bounded optimal value functions. It has enough structure, however, that it is still possible to demonstrate uniform geometric convergence of the method of successive approximations for certain starting functions. The structure of the model includes that found in a large number of Markov decision processes, including most of those studied in the literature on stochastic service systems in which monotonic, or critical-number, policies are optimal. This structure makes it possible to transform the model, by means of a shift function, into an equivalent model satisfying the conditions of Section 2. We illustrate the applicability of the model by some examples from queueing and inventory control.

We shall use the following two conditions throughout this section.

Condition 3.1. $\rho := \sup_{f \in F} \| P(f) \| < 1$

Condition 3.2. $M := \sup_{f \in F} \| r^{+}(f) \| < \infty$

Note that Conditions 3.1 and 3.2 are exactly the conditions of the essentially negative case.

In this section we shall be interested in the Banach space $\mathcal{V}(V)$, where $V: S \to \underline{R}$ is a given non-zero reference function. We shall use the following condition on V.

Condition 3.3. $\| UV - V \| < \infty$.

An upper bound on V^* is a natural candidate for a reference function, since in many applications (see examples later in this section) it is easy to compute such an upper bound. Since Conditions 3.1 and 3.2 imply that $V^* \leq M_1 = M(1-\rho)^{-1} < \infty$, it also makes sense to confine our attention to reference functions $\bar{V} \geq V^*$ such that $\bar{V} \leq M_1$.

Theorem 3.1. Let $V = \bar{V}$, where $V^* \leq \bar{V} \leq M_1$. Assume Conditions 3.1 - 3.3. Then $V^* \in W(\bar{V})$ and is the unique solution in $W(\bar{V})$ to the optimality equation

$$v = Uv$$

and, for any $V_o \in W(\bar{V})$, $\| V^* - U^n V_o \| \leq \rho^n \| V^* - V_o \| < \infty$, so that $U^n V_o$ converges to V^* uniformly and geometrically.

Proof. Condition 3.1 implies (ii) of Lemma 1.1. It remains to verify (i) of Lemma 1.1: $V^* \in W(\bar{V})$.

To this end, we first claim that it suffices to show pointwise convergence of successive approximations with $\bar{V}$ as starting function:

$$U^n \bar{V} \to V^* \tag{3.1}$$

To see this, note that (ii) of Lemma 1.1. and Condition 3.3 imply that ($s \in S$)

$$\begin{aligned} U^n \bar{V}(s) - \bar{V}(s) &\leq \sum_{k=1}^{n-1} \| U^{k+1}\bar{V} - U^k \bar{V} \| \\ &\leq \sum_{k=1}^{n-1} \rho^k \| U\bar{V} - \bar{V} \| \\ &\leq (1-\rho)^{-1} \| U\bar{V} - \bar{V} \| < \infty \end{aligned}$$

Letting $n \to \infty$ and assuming (3.1), we conclude that ($s \in S$)

$$V^*(s) - \bar{V}(s) \leq (1-\rho)^{-1} \| U\bar{V} - \bar{V} \| < \infty .$$

Reversing the roles of $U^n\bar{V}(s)$ and $\bar{V}(s)$ yields the same uniform upper bound for $\bar{V}(s) - V^*(s)$, so that

$$\| V^* - \bar{V} \| \leq (1-\rho)^{-1} \| U\bar{V} - \bar{V} \| < \infty,$$

that is, $V^* \in W(\bar{V})$.

Thus it remains to prove (3.1). Define $V_\infty = \lim_{n\to\infty} U^n\bar{V}$. Since $\bar{V} \geq V^*$, $U^n\bar{V} \geq U^n V^* = V^*$, so that $V_\infty \geq V^*$. Hence it suffices to show that $\overline{\lim}_{n\to\infty} U^n\bar{V} \leq V^*$.

To this end, we first show that Conditions 3.1 - 3.3 imply that $V_\infty = \lim_{n\to\infty} U^n\bar{V}$ and that V_∞ is a fixed point of U. Although a direct proof is possible, we shall instead prove these facts by means of a shift transformation, which converts the problem into an equivalent problem satisfying the conditions of Model I. This shift transformation is of independent interest. In

particular, it has an interesting economic interpretation, which we shall discuss later in this section.

For each decision rule f, define $\tilde{r}(f) := r(f) + P(f)\bar{V} - \bar{V}$.

Consider the Markov decision model $(S,A,D,\tilde{r},p)$. Condition 3.3 implies that:

$$\| \sup_{f\in F} \tilde{r}(f) \| = \| \sup_{f\in F}\{r(f) + P(f)\bar{V}\} - \bar{V} \|$$
$$= \| U\bar{V} - \bar{V} \| < \infty,$$

so that Condition 2.2 holds. Condition 3.1 is identical to Condition 2.1. Define the operator $\tilde{U}$, where it is well defined, by

$$\tilde{U}v := \sup_{f\in F}\{ \tilde{r}(f) + P(f)v \}.$$

Define the infinite-horizon optimal value function $\tilde{V}^*$ for the transformed model in the obvious way (cf. Section 1). Then

$$\tilde{V}^* = \tilde{U}\tilde{V}^*.$$

Moreover, since the transformed model satisfies conditions 2.1 and 2.2, it follows from Theorem 2.1 that $\tilde{V}^*$ is the unique bounded fixed point of $\tilde{U}$ and that

$$\tilde{V}^* = \lim_{n\to\infty} \tilde{U}^n 0. \tag{3.2}$$

Now observe that

$$\tilde{U}0 = \sup_{f\in F} \tilde{r}(f)$$
$$= \sup_{f\in F}\{ r(f) + P(f)\bar{V} \} - \bar{V}$$
$$= U\bar{V} - \bar{V},$$

and by induction on n,

$$\tilde{U}^n 0 = U^n\bar{V} - \bar{V}, \quad n \geq 1. \tag{3.3}$$

From (3.2) and (3.3) we conclude that

$$V_\infty = \lim_{n\to\infty} U^n\bar{V} = \tilde{V}^* + \bar{V}.$$

Hence

$$\begin{aligned} UV_\infty &= U(V^* + \bar{V}) \\ &= \sup_{f\in F} \{r(f) + P(f)\ (\tilde{V}^* + \bar{V})\ \} \\ &\le \sup_{f\in F} \{r(f) + P(f)\bar{V} - \bar{V} + P(f)\tilde{V}^*\ \} + \bar{V} \\ &= \sup_{f\in F} \{r(f) + P(f)\tilde{V}^*\ \} + \bar{V} \\ &= \tilde{U}\tilde{V}^* + \bar{V} \\ &= \tilde{V}^* + \bar{V} \\ &= V_\infty\ , \end{aligned}$$

that is, V_∞ is a fixed point of U, the desired result.

Let $\varepsilon>0$ be given. Choose a stationary policy f such that

$$V_\infty = UV_\infty \le L(f)\ V_\infty + \varepsilon\ . \tag{3.4}$$

(See Remark 3.1 below.) Iterating on L(f) and using Condition 3.1 , we have

$$\begin{aligned} V_\infty &\le (L(f))^n V_\infty + \sum_{k=0}^{n-1} \rho^k \varepsilon \\ &= (L(f))^n 0 + (P(f))^n V_\infty + \sum_{k=0}^{n-1} \rho^k \varepsilon \end{aligned}$$

so that

$$\begin{aligned} V_\infty &\le V(f) + \overline{\lim_{n\to\infty}}\ (P(f))^n V_\infty + \varepsilon(1-\rho)^{-1} \\ &\le V^* + \overline{\lim_{n\to\infty}} (P(f))^n V_\infty + \varepsilon(1-\rho)^{-1}. \end{aligned}$$

From Conditions 3.1 and 3.2 and the hypothesis that $\bar{V} \le M_1$ it follows that $V_\infty = \lim_{n\to\infty} U^n\bar{V} \le M_1$, so that

$$\overline{\lim_{n\to\infty}}\ (P(f))^n V_\infty \le \lim_{n\to\infty} \rho^n M_1 = 0.$$

Hence, since ε was arbitrary, we conclude that $V_\infty \le V^*$, the desired result.

Remark 3.1. The existence of a policy f satisfying (3.4) follows from the existence in general of ε - optimizing decision rules, which is a mild regularity condition, apparently satisfied in all practical problems. It holds, for example, when the state space is countable. For general state space, where

it is customary to require decision rules to be measurable functions in order for the relevant stochastic processes and integrals to be well defined, there may be difficulties in applying the condition to certain functions v, such as $v = V_\infty$, which may not themselves be measurable (cf. Blackwell [3], Strauch [34], Hinderer [11]). There is no problem, for example, if the action space is countable or if continuity - compactness conditions are satisfied by (S,A,D,r,p) (see, e.g., Schäl [24]). Alternatively, one may enlarge the class of admissible decision rules to include all universally measurable functions $f: S \to A$ such that $f(s) \in D(s)$, $s \in S$, and require that r, p, and v also be universally measurable (cf. Shreve and Bertsekas [27]).

Remark 3.2. It can easily be shown that Condition 3.3 and (3.1) together are necessary as well as sufficient for (i) of Lemma 1.1. In other words:

$$V^* \in \mathcal{V}(V) \quad \text{iff} \quad \| UV - V \| < \infty \quad \text{and} \quad U^n V \to V^*.$$

We now give several examples of applications of Model II to problems in control of queues, followed by some suggestions for a general approach to the solution of such problems. Finally, we show how Model II can also be applied to certain periodic-review inventory models.

Example 1°. Control of Arrivals.

Again we use the model of Johansen and Stidham [12] for control of arrivals to a stochastic input-output system as a vehicle for illustrating the verification and application of our conditions. (See Section 2 for a detailed introduction to the model.) In the present context we are interested in the case of continuous, rather than lump-sum, charging of the holding cost. To be specific, the system is observed at arrival points, the state s is the quantity in the system found by an arrival, the action a indicates acceptance (a = 1) or rejection (a= 0) of the potential input of the arrival, the return function r(s,a) is given by:

$$r(s,a) = ar - E[\int_0^T e^{-\alpha\tau} h((s + aS - N(\tau))^+) \, d\tau \,],$$

and the transition probability by

$$p(s,a;B) = E[e^{-\alpha T} \underset{\sim}{1} ((s + aS - N(T))^+ \in B)].$$

Here $h(\cdot)$ gives the rate (per unit time) at which holding cost is incurred, as a function of the quantity in the system. We assume that $h(\cdot)$ is a non-negative, non-decreasing, convex mapping from S into R.

Thus the model differs from the one considered in Section 2 only with respect to the return function. In the model of Section 2, there is a waiting cost, C(s), associated with an arriving customer who joins when the state is s. This cost might represent the expected discounted cost that the customer will incur during the entire time he spends waiting in the system. Note that it is a cost associated only with the joining customer, but that it reflects time spent waiting after as well as until the next arrival point. By contrast, in the present model the return function includes costs associated with all customers in the system, but only until the next arrival. The relation between the two charging schemes is given by

$$C(s) = E[\int_0^\infty e^{-\alpha\tau} \, [h((s + S - N(\tau))^+) - h((s - N(\tau))^+] \, d\tau]. \tag{3.5}$$

It follows from the assumption that $h(\cdot)$ is non-decreasing and convex that $C(\cdot)$ defined by (3.5) is non-negative and non-decreasing, as required in the model of Section 2.

(For further discussion and economic interpretation of the two charging schemes, see Johansen and Stidham [12].)

The infinite-horizon optimality equation, $V^* = UV^*$, for the problem with continuous charging is given explicitly by ($s \geq 0$)

$$V^*(s) = \max_{a=0,1} \{ar - E[\int_0^T e^{-\alpha\tau} h((s + aS - N(\tau))^+) \, d\tau]$$

$$+ E[e^{-\alpha T} V^*((s + aS - N(T))^+)] \, .$$

It is easily verified that Conditions 3.1 and 3.2 hold. In fact, $V^* < M_1 = r(1-\rho)^{-1} < \infty$, where $\rho = E[e^{-\alpha T}] < 1$.

Define $\bar{V} : S \to \underline{R}$ by

$$\bar{V}(s) := r(1-\rho)^{-1} - E[\int_0^\infty e^{-\alpha\tau} h((s-N(\tau))^+) \, d\tau] \qquad (s \in S) \quad (3.7)$$

Theorem 3.2. Consider the arrival-control model with continuous charging of holding cost and with $\bar{V}$ defined by (3.7). Then $V^* \leq \bar{V} < M_1 < \infty$, $V^* \in \mathcal{W}(\bar{V})$ and is the unique solution in $\mathcal{W}(\bar{V})$ to the optimality equation (3.6). Moreover, for any $V_0 \in \mathcal{W}(\bar{V})$, $\| V^* - U^n V_0 \| \leq \rho^n \| V^* - V_0 \| < \infty$, so that $U^n V_0$ converges to V^* uniformly and geometrically.

Proof. Let π be an arbitrary policy, noting that

$$V^\pi(s) = E_s^\pi [\sum_{t=0}^\infty r(X_t, A_t)] \qquad (3.8)$$

$$= E_s^\pi [\sum_{t=0}^\infty r A_t] - E_s^\pi [\sum_{t=0}^\infty E[\int_0^T e^{-\alpha\tau} h((X_t + A_t S - N(\tau))^+) d\tau | X_t, A_t]]$$

The first term on the right-hand side of (3.8) is maximized by the (stationary) policy that accepts all customers, whereas the second term is maximized by the policy that minimizes holding costs, namely, the (stationary) policy f_r that rejects all customers ($f_r(s) = 0$, for all $s \in S$). Note that

$$V^{f_r}(s) = -E[\int_0^\infty e^{-\alpha\tau} h((s-N(\tau))^+) d\tau] \qquad (s \in S) \qquad (3.9)$$

since under f_r no rewards are earned, but all the quantity s currently in the system must be processed and will incur holding costs until it is processed. It follows then from (3.7), (3.8), and (3.9) that

$$V^\pi \leq M_1 + V^{f_r} = \bar{V} \qquad (3.10)$$

since π was arbitrary and $V^{f_r} \leq 0$, we conclude that $V^* \leq \bar{V} \leq M_1 < \infty$.

To complete the proof it suffices to show that Condition 3.3 holds, so that Theorem 3.1 applies. To this end, first observe that V^{fr} satisfies the functional equation.

$$V^{fr}(s) = -E[\int_0^T e^{-\alpha\tau} h((s-N(\tau))^+)d\tau] + E[e^{-\alpha T} V^{fr}((s-N(T))^+)] \quad (s \in S) \qquad (3.11)$$

(This follows from (1.1) and (3.9).) Note also that (3.5) and (3.9) imply that

$$C(s) = -E[V^{fr}(s + S)] + V^{fr}(s). \qquad (3.12)$$

Now, using (3.10), (3.11), and (3.12), we can write

$$\begin{aligned} U\bar{V}(s) - \bar{V}(s) &= \max_{a=0,1} \{ar - E[\int_0^T e^{-\alpha\tau} h((s+aS-N(\tau))^+)d\tau] \\ &\quad + E[e^{-\alpha T}\bar{V}((s + aS - N(T))^+]\} - \bar{V}(s) \\ &= \max_{a=0,1} \{ar - E[\int_0^T e^{-\alpha\tau} h((s+aS - N(\tau))^+)d\tau] \\ &\quad + E[e^{-\alpha T}V^{fr}((s + aS -N(T))^+] - V^{fr}(s) - (1-\rho)M_1 \\ &= \max_{a=0,1} \{ar - E[V^{fr}(s + aS)] \} - V^{fr}(s) - r \\ &= \max \{r - C(s), 0\} - r. \end{aligned} \qquad (3.13)$$

Since $C(\cdot)$ is non-decreasing, it follows from (3.13) that

$$-r \leq U\bar{V}(s) - \bar{V}(s) \leq 0 \qquad (3.14)$$

and hence $\| U\bar{V} - \bar{V} \| < r < \infty$, so that Condition 3.3 holds. This completes the proof of the theorem.

Remark 3.3. Since $U\bar{V} \leq \bar{V}$ (by (3.13)), it follows by induction from the monotonicity of U that $U^n\bar{V} \leq U^{n-1}\bar{V}$, for all $n \geq 1$, so that the convergence of $U^n\bar{V}$ to V* is monotonically decreasing.

Corollary 3.3. Consider the arrival-control model with continuous charging of holding cost. Suppose $V = V^{fr} \leq V^*$. Then $V^* \in \mathcal{U}(V)$ and is the unique solution in $\mathcal{U}(V)$ to the optimality equation (3.5). Moreover, for any $V_o \in \mathcal{U}(V)$, $\| V^* - U^n V_o \| \leq \rho^n \| V^* - V_o \| < \infty$, so that $U^n V_o$ converges to V* uniformly and geometrically.

Proof. A direct consequence of Theorem 3.2 since (3.10) implies that $\| \bar{V} - V^{f_r} \| = M_1 < \infty$, where $\bar{V}$ is defined by (3.7).

Remark 3.4. Since $UV^{f_r} \geq L(f_r)V^{f_r} = V^{f_r}$, it follows by induction from the monotonicity of U that $U^nV \geq U^{n-1}V$, for all $n \geq 1$, when $V = V^{f_r}$. Hence the convergence of U^nV to V* is monotonically increasing. In fact, of course, convergence of U^nV to V* is monotonically increasing whenever V is the infinite-horizon value function for a particular policy. This observation is true in general of "successive approximations in policy space" and goes back at least to Bellman [1].

Remark 3.5. Define a new return function $\check{r}$ by $\check{r}(s,a) := r(s,a) + E[e^{-\alpha T}V^{f_r}((s + aS - N(T))^+)] - V^{f_r}(s)$ $(s \in S, a = 0,1)$ or, equivalently, $\check{r}(f) := r(f) + P(f)V^{f_r} - V^{f_r}$, for each decision rule f. In fact, the proof of Theorem 3.2 shows that

$$\check{r}(s,a) = a(r - C(s)) \qquad (s \in S, a = 0,1)$$

so that the Markov decision model $(S,A,D,\check{r},p)$ has the structure of the arrival-control problem with lump-sum charging of waiting costs, which was studied in Section 2. The effect of this shift transformation is to subtract V^{f_r} from the value function for each policy and hence from the optimal value functions for both finite and infinite horizons. In economic terms, $-V^{f_r}(s)$ is just the expected discounted cost of holding the quantity s until it is processed and hence represents an unavoidable, or fixed, cost that must be incurred no matter what policy is followed. The shift transformation thus can be interpreted as a removal of these fixed costs, so that only those costs that vary with the policy - namely, those associated with the current and future customers - are included in the value functions. That such a transformation should lead to an equivalent decision problem is therefore plausible on intuitive grounds.

This shift transformation was first proposed for an arrival-control problem in Lippman and Stidham [16]. In that context, however, it was used for a different purpose, namely, facilitating the comparison of an optimal policy with the (equilibrium) policy followed when each customer acts to maximize his own expected discounted net benefit: $a(r - C(s))$. (Note that an equilibrium policy is myopic with respect to the transformed model $(S,A,D,\tilde{r},p)$.)

An equilibrium policy can also be used to generate an alternative reference function V for the model with continuous charging. Denote by f_e the (stationary) equilibrium policy; that is, f_e accepts a customer ($f_e = 1$) in state s iff $r \geq C(s)$. Since the optimal value function, $\tilde{V}^*$, for the transformed model $(S,A,D,\tilde{r},p)$ satisfies (2.3), it follows that an optimal policy accepts a customer iff $r \geq C(s) + E[e^{-\alpha T}(\tilde{V}^*((s - N(T))^+) - \tilde{V}^*((s + S - N(T))^+))]$. It can easily be shown (see Johansen and Stidham [12]) that $\tilde{V}^*(\cdot)$ is non-increasing. Hence an optimal policy accepts a customer in state s only if the equilibrium policy f_e accepts in s (cf. also Lippman and Stidham [16] and the references cited therein for other instances of this phenomenon). This property can be used to prove

Theorem 3.4. Consider the arrival-control model with continuous charging of holding cost. Suppose $V = V^{f_e} \leq V^*$. Then $V^* \in \mathcal{W}(V)$ and is the unique solution in $\mathcal{W}(V)$ to the optimality equation (3.6). Moreover, for any $V_o \in \mathcal{W}(V)$, $\| V^* - U^n V_o \| \leq \rho^n \| V^* - V_o \| < \infty$, so that $U^n V_o$ converges to V^* uniformly and geometrically.

Proof. We shall verify that $U^n V \to V^*$ and that Condition 3.3 holds. The remainder of the proof then parallels that of Theorem 3.1. (See Remark 3.2.)

Let f^* denote an optimal policy. Since f^* accepts only if f_e accepts, it follows that for each $n \geq 1$ X_n is stochastically smaller under f^* than under f_e, given $X_o = s$. Thus, since $V^{f_e}(\cdot)$ is non-increasing, we have (cf. Lemma 2.3)

$$(P(f^*))^n V^{f_e} = E^{f^*}[V^{f_e}(X_n)]$$

$$\geq E^{f_e}[V^{f_e}(X_n)] = (P(f_e))^n V^{f_e} \tag{3.15}$$

But

$$V^{f_e} = L(f_e)V^{f_e} = (L(f_e))^n V^{f_e}$$

$$= (L(f_e))^n 0 + (P(f_e))^n V^{f_e}$$

and $(L(f_e))^n 0 \to V^{f_e}$, as $n \to \infty$, by (1.2), so that $(P(f_e))^n V^{f_e} \to 0$, as $n \to \infty$.

Using this fact together with (3.15), we conclude that

$$\overline{\lim_{n\to\infty}} (P(f^*))^n V^{f_e} \geq \overline{\lim_{n\to\infty}} (P(f_e))^n V^{f_e} = 0. \tag{3.16}$$

Now it follows from (3.16) and Theorem 3.5 of [32] that $U^n V^{f_e} \to V^*$.

To verify Condition 3.3, first observe that ($s \in S$)

$$UV^{f_e}(s) = \max_{a=0,1} \{ar - E[\int_0^T e^{-\alpha\tau} h((s+aS-N(\tau))^+)d\tau] + E[e^{-\alpha T} V^{f_e}((s+aS-N(T))^+)]\}$$

$$= \max\{r + E[W(s+S)], W(s)\}$$

$$= W(s) + (r - w(s)) \mathbb{1}(w(s) \leq r) \tag{3.17}$$

where.

$$W(s) := -E[\int_0^T e^{-\alpha\tau} h((s-N(\tau))^+) dt] + E[e^{-\alpha T} V^{f_e}((s-N(T))^+)]$$

$$w(s) := W(s) - E[W(s+S)].$$

On the other hand,

$$V^{f_e}(s) = L(f_e) V^{f_e}(s)$$

$$= (r+E[W(s+S)]) \mathbb{1}(C(s) \leq r) + W(s) \mathbb{1}(C(s) > r)$$

$$= W(s) + (r - w(s)) \mathbb{1}(C(s) \leq r) \tag{3.18}$$

It follows from (3.17) and (3.18) that

$$0 \leq UV^{f_e}(s) - V^{f_e}(s)$$

$$= (r-w(s)) \mathbb{1}(w(s) \leq r < C(s)) + (w(s)-r) \mathbb{1}(C(s) \leq r < w(s)). \tag{3.19}$$

It remains to show that $UV^{f_e} - V^{f_e}$ is uniformly bounded above. To this end, first observe that

$$V^{f_r}(s) \leq V^{f_e}(s) \leq M_1 + V^{f_r}(s). \tag{3.20}$$

The right-hand inequality follows from (3.10). The left-hand inequality can be verified as follows. First, note that $L(f_r)V^{f_e}(s) = \mathbb{1}(r \geq C(s))\ (r+E[V^{f_r}(s+S)])$

$+ \underset{\sim}{1}(r < C(s))\ V^{f_r}(s) = V^{f_r}(s) + \underset{\sim}{1}(r > C(s))\ \ (r - C(s))^+ \geq V^{f_r}(s)$. (Here we have used (3.12).) Hence, iterating and using the monotonicity of the operator $L(f_e)$, we have $(n \geq 1)$

$$V^{f_r} \leq (L(f_e))^n V^{f_r} = (L(f_e))^n 0 + (P(f_e))^n V^{f_r}.$$

Letting $n \to \infty$ and using the fact that $V^{f_r} \leq 0$, we have

$$V^{f_r} \leq V^{f_e} + \overline{\lim_{n\to\infty}} (P(f_e))^n V^{f_r} \leq V^{f_e}.$$

Now it follows from (3.11) and the definition of W(s) that

$$W(s) = V^{f_r}(s) + E[e^{-\alpha T}(V^{f_e}((s - N(T))^+) - V^{f_r}((s - N(T))^+))].$$

Hence, from (3.20) we obtain

$$V^{f_r}(s) \leq W(s) \leq V^{f_r}(s) + \rho M_1,$$

which imples (using (3.12)) that

$$-\rho M_1 + C(s) \leq w(s) \leq C(s) + \rho M_1.$$

Using these inequalities together with (3.19), we find that

$$\begin{aligned} 0 \leq UV^{f_e}(s) - V^{f_e}(s) &\leq (r - C(s) + \rho M_1)\ \underset{\sim}{1}\ (w(s) \leq r < C(s)) \\ &\quad + (C(s) - r + \rho M_1)\ \underset{\sim}{1}\ (C(s) \leq r < w(s)) \\ &\leq \rho M_1 [\underset{\sim}{1}\ (w(s) < r < C(s)) + \underset{\sim}{1}\ (C(s) \leq r < w(s))] \\ &\leq \rho M_1 < \infty, \end{aligned}$$

thus verifying Condition 3.3.

Remark 3.6. The proof of Theorem 3.4 could be simplified considerably if one could assert that w(s) is non-decreasing and $w(s) \geq C(s)$. The first property says that, if one is free to choose the action in the current period but must follow f_e thereafter, then the optimal action is non-increasing in s. The second property says that, under the same circumstances, the optimal action will be to reject (a = 0) whenever the equilibrium policy rejects ($f_e(s) = 0$), and perhaps in other states as well. (Analogous properties hold when one considers an optimal rather than equilibrium policy; see [12].) However, we have not been able to find proofs for either property and conjecture that they do not

hold in general. Intuitively, if one must follow an equilibrium policy after the current period, then it might be optimal to accept in the current period even if an equilibrium policy would reject, since the resulting state at the next arrival might then be large enough to force the next customer to balk rather than join, which in turn makes it possible for the customer after that to join rather than balk. The net increase in total welfare could be positive in some cases.

Example 2°. Control of the Service Rate

As a further illustration of how our theorems can be applied to problems in queueing control, we consider an M/M/1 queue with variable service rate. (See Stidham and Prabhu [33], Sobel [28], or Crabill, Gross, and Magazine [4] for detailed discussion of this model and relevant references.) Let the arrival rate $\lambda > 0$ be fixed. The service rate μ can be chosen from the interval $[0, \bar{\mu}]$. Whenever service rate μ is in effect, the system incurs a cost $c(\mu)$ per unit time, where $c(\cdot)$ is non-decreasing and continuous with $c(0) = 0$. A holding cost is incurred at rate $h(i)$ whenever i customers are in the system, $i \geq 0$, where $h(\cdot)$ is non-negative and non-decreasing. Costs are discounted continuously at rate $\alpha > 0$.

To construct a Markov decision model for this problem, we use the "new device" of Lippman [15] and observe the system at points of arrival (occurring at rate λ), service completion (occurring at rate μ), and null events (occurring at rate $\bar{\mu}-\mu$). The time between observation points (stages) has exponential distribution with parameter $\Lambda = \lambda + \bar{\mu}$. The system is said to be in state i (where i is a non-negative integer) whenever there are i customers present. The action taken at an observation point is the service rate μ, which remains in effect until the next observation point. Thus, $D(i) = [0, \bar{\mu}]$, $i \geq 1$, and $D(0) = \{0\}$. The one-stage return function $r(i,\mu)$ is given by

$$r(i, \mu) = (\alpha + \Lambda)^{-1} [-c(\mu) - h(i)]$$

and the transition probability measure is determined by the discrete (discounted) transition probabilities,

$$p_{ij}(\mu) := \Pr\{X_{t+1} = j \mid X_t = i, A_t = \mu\}$$
$$= (\alpha + \Lambda)^{-1} [\lambda 1(j=i+1) + \mu 1(j=i-1) + (\bar{\mu}-\mu) 1(j=i)].$$

Conditions 3.1 and 3.2 are satisfied, with $\rho = \Lambda(\Lambda + \alpha)^{-1}$ and $M=0$. Define $V^*(i)$ $(i=0,1,\ldots)$ as the optimal value function for the infinite-horizon problem. V^* satisfies the optimality equations

$$V^*(i) = (\Lambda+\alpha)^{-1} \max_{0 \le \mu \le \bar{\mu}} \{-c(\mu) - h(i) + \lambda V^*(i+1) + \mu V^*(i-1) + (\bar{\mu}-\mu)V^*(i)\},$$
$$i \ge 1 \qquad (3.21)$$
$$V^*(0) = (\Lambda+\alpha)^{-1} \{-h(0) + \lambda V^*(1) + \bar{\mu} V^*(0)\}$$

Define the stationary policy g by $g(i) = \bar{\mu}$, $i = 1,2,\ldots$. We call g the full-service policy. Among all policies, g obviously minimizes the infinite-horizon expected discounted holding cost. Now define the function $\bar{V}: S \to \underline{R}$ by

$$\bar{V}(i) := -(\alpha + \Lambda)^{-1} E_i^g[\sum_{t=0}^{\infty} h(X_t)], \quad i \ge 0 \qquad (3.22)$$

Theorem 3.5. Consider the M/M/1 service-rate-control model, with $\bar{V}$ defined by (3.22). Then $V^* \le \bar{V} < 0$, $V^* \in W(\bar{V})$ and is the unique solution in $W(\bar{V})$ to the optimality equations (3.21). Moreover, for any $V_0 \in W(\bar{V})$, $\|V^* - U^n V_0\| \le \rho^n \|V^* - V_0\| < \infty$, so that $U^n V_0$ converges uniformly and geometrically to V^*.

Proof. Let π be an arbitrary policy and note that

$$V^\pi(i) = E_i^\pi[\sum_{t=0}^{\infty} r(X_t, A_t)] \qquad (3.23)$$
$$= -(\alpha + \Lambda)^{-1} E_i^\pi[\sum_{t=0}^{\infty} c(A_t)] - (\alpha+\Lambda)^{-1} E_i^\pi[\sum_{t=0}^{\infty} h(X_t)], \quad i \ge 0$$

The first term on the right-hand side of (3.23) is non-positive and the second term is maximized by the full-service policy g, since g minimizes the infinite-horizon expected discounted holding costs. Therefore, $V^\pi \le \bar{V}$. Since π was arbitrary we conclude that

$$V^* \le \bar{V} < 0.$$

To complete the proof, it suffices to show that Condition 3.3 holds, so that Theorem 3.1 may be applied. To this end, first observe that $\bar{V}$ satisfies the functional equations

$$\bar{V}(i) = (\alpha+\Lambda)^{-1}\{-h(i) + \lambda\bar{V}(i+1) + \bar{\mu}\bar{V}(i-1)\}, \quad i \geq 1$$
$$\bar{V}(0) = (\alpha+\Lambda)^{-1}\{-h(0) + \lambda\bar{V}(1) + \bar{\mu}\bar{V}(0)\}$$

Therefore, $U\bar{V}(0) - \bar{V}(0) = 0$ and, for $i \geq 1$,

$$\begin{aligned} U\bar{V}(i) - \bar{V}(i) &= (\alpha+\Lambda)^{-1} \left[\max_{0\leq\mu\leq\bar{\mu}} \{-c(\mu)-h(i)+ \lambda\bar{V}(i+1) + \mu\bar{V}(i-1) + (\bar{\mu}-\mu)\bar{V}(i) \right. \\ &\qquad \left. - \{-h(i) + \lambda\bar{V}(i+1) + \bar{\mu}\bar{V}(i-1)\}\right] \\ &\leq (\alpha+\Lambda)^{-1} \max_{0\leq\mu\leq\bar{\mu}} \{(\bar{\mu}-\mu)\,(\bar{V}(i) - \bar{V}(i-1))\} \end{aligned}$$

Since $\bar{V}(i) \leq \bar{V}(i-1)$, for all $i \geq 1$, the quantity in brackets is maximized by setting $\mu = \bar{\mu}$. Hence,

$$U\bar{V}(i) - \bar{V}(i) \leq 0, \qquad i \geq 1. \tag{3.24}$$

On the other hand,

$$U\bar{V}(i) - \bar{V}(i) \geq L(g)\,\bar{V}(i) - \bar{V}(i) = -c(\bar{\mu}), \qquad i \geq 1.$$

Therefore, Condition 3.3 holds and the theorem is proved.

Remark 3.7. Since $U\bar{V} \leq \bar{V}$ (by (3.24)), it follows by induction using the monotonicity of U that $U^n\bar{V} \leq U^{n-1}\bar{V}$, for all $n \geq 1$, so that the convergence of $U^n\bar{V}$ to V^* is monotonically decreasing.

Corollary 3.6. Consider the M/M/1 service-rate-control model. Suppose $V=V^g \leq V^*$. Then $V^* \in \mathcal{W}(V)$ and is the unique solution in $\mathcal{W}(V)$ to the optimality equations (3.21). Moreover, for any $V_0 \in \mathcal{W}(V)$, $\|V^* - U^nV_0\| \leq \rho^n\|V^* - V_0\| < \infty$, so that U^nV_0 converges uniformly and geometnically to V^*.

Proof. It follows from (3.22) and (3.23) that

$$\bar{V}(i) \geq V^g(i) \geq \alpha^{-1}\, c(\bar{\mu}) + \bar{V}(i), \qquad i \geq 0.$$

Thus the corollary follows immediately from Theorem 3.5.

Remark 3.8. Since $UV^g \geq L(g)\,V^g = V^g$, it follows by induction using the monotonicity of U that $U^nV^g \geq U^{n-1}V^g$, for all $n \geq 1$, so that the convergence of

$U^n V^g$ to V^* is monotonically increasing.

Remark 3.9. Define a new return function $\tilde{r}$ by ($i = 0,1,\ldots$, $0 \le \mu \le \bar{\mu}$)

$$\tilde{r}(i,\mu) := r(i,\mu) + \sum_{j=0}^{\infty} p_{ij}(\mu)\ V^g(j) - V^g(i),$$

or, equivalently,

$$\tilde{r}(f) := r(f) + P(f)\ V^g - V^g$$

for each decision rule $f \in F$. By an argument similar to that used in the proof of Theorem 3.5, it follows that

$$\tilde{r}(i,\mu) = (\alpha+\Lambda)^{-1}\ [c(\bar{\mu}) - c(\mu) - (\bar{\mu} - \mu)\ (V^g(i-1) - V^g(i))],$$

$$i \ge 1, \quad \mu \in [0,\bar{\mu}] \tag{3.25}$$

$$\tilde{r}(0,0) = 0.$$

From (3.25) we see immediately that

$$0 \le \max_{0 \le \mu \le \bar{\mu}} \tilde{r}(i,\mu) \le (\alpha+\Lambda)^{-1}\ c(\bar{\mu}),$$

so that the Markov decision model $(S,A,D,\tilde{r},p)$ satisfies the conditions of Model I studied in Section 2. The effect of this shift transformation is to subtract V^g from the value function for each policy and hence from the optimal value function for both finite and infinite horizons. Thus, the value functions for all policies are measured relative to that of an extremal (in this case, full-service) policy. Like the policy that rejects all customers in the arrival-control problem, this full-service policy minimizes holding costs among all policies. Thus, the holding costs under policy g -- that is, $-\bar{V}(i)$ as defined by (3.22) -- are unavoidable fixed costs that must be incurred no matter what policy we follow. (Other policies may incur larger holding costs, of course.) Our results say that the original service-rate-control model is equivalent to one in which these fixed costs have been eliminated, that is, a model with $\tilde{r}(i,\mu)$ as its one-stage return function.

Examination of (3.25) reveals that $\tilde{r}(i,\mu)$ can be interpreted as the net savings from taking action μ rather than $\bar{\mu}$ in the current period, if policy g

is to be followed in all future periods. If the maximum net savings is zero, that is, if

$$(\bar{\mu}-\mu)^{-1}[c(\bar{\mu}) - c(\mu)] < V^g(i-1) - V^g(i), \text{ for all } \quad i \geq 1,$$

then a full-service policy is optimal (cf. Sobel [29], Schleef [25], Sobel and Winston [30]). Thus, in addition to providing a convenient vehicle for demonstrating uniform, geometric convergence of successive approximations, the transformed model with V^g as shift function is in some sense the "proper setting" in which to investigate the form of an optimal policy. We have already made a similar observation in the context of the arrival-control model, and we believe the observation to be valid in the majority of control models for stochastic service systems and related systems.

Remark 3.10. A close examination of the proofs of Theorems 3.2 and 3.5 will reveal that in both cases the proof of Condition 3.3 for $V = \bar{V}$ hinges on the fact that the function

$$\begin{aligned} h(s,a) &:= L(f_a)\,\bar{V}(s) \\ &= r(s,a) + \int_S p(s,a,ds')\,\bar{V}(s') \end{aligned}$$

is supermodular [36], [35], [33] in (s,a), where f_a is the stationary policy that always takes action a ($f_a(s) = a$, for all $s \in S$). This property, together with the facts that $\bar{V}(s)$ is nonincreasing in $s \in S$, S contains a minimal element 0, and $\bar{V}$ (in both examples) is related to the value function for an extremal policy, leads directly to uniform upper and lower bounds on $U\bar{V}(s) - \bar{V}(s)$ and thence to Condition 3.3. Supermodularity is extensively used as a device for proving monotonicity of optimal control policies [26], [28], [31], [12]. We expect that it could also be used as the basis for a general model of queueing-control problems for which methods like those used in Theorem 3.1 could be used to demonstrate uniform geometric convergence of successive approximations.

Example 3^{o}. Inventory Control.

This example will be used to give three variants of the classical single-product inventory problem with periodic review, as described, e.g., by Scarf [23]. We shall allow for unbounded returns and not require convexity conditions for the cost or reward structure. Let us first describe the general model.

An inventory system is observed at discrete points in time, say the beginning of each month. At these points in time the state of the system is defined as the available inventory. This inventory may be negative as well as positive, so backlogging is allowed. Let us represent the state space by $S = R = (-\infty, \infty)$. If the state $s \in S$ is observed at time $t \in \{0,1,2,\ldots\}$, then a positive amount $a \in R^{+}$ can be ordered. Delivery is immediate. The ordering cost $c(a)$ is non-decreasing in $a \geq 0$. A holding cost $h(s)$ is incurred at the beginning of a period and is non-decreasing as a function of the inventory $s \geq 0$. If the observed inventory s at the beginning of a period is negative, then a shortage cost $p(-s)$ is incurred. We assume $p(\ )$ to be a non-decreasing function of the amount of the shortage. Finally, we assume that the demands in successive periods are i.i.d. random variables with probability distribution function $\phi(\xi)$, such that the expected demand in each period is finite, i.e.,

$$\int_{0}^{\infty} \xi \, d\phi(\xi) = : M_1 < \infty. \tag{3.26}$$

Costs and rewards are discounted. The one-period discount factor is $\rho < 1$, so that costs incurred in period t are weighted with the factor ρ^{t}. As was the case with the queueing-control examples, the goal is to maximize the total expected discounted returns (equivalently: minimize the total expected discounted costs) over an infinite horizon, and to find a (stationary) policy for which this maximum is (approximately) achieved.

In the notation of our basic decision model, the one-stage return function is given by

$$r(s,a) = \begin{cases} -h(s) - c(a), & \text{for } s \geq 0,\ a \geq 0 \\ -p(-s) - c(a), & \text{for } s < 0,\ a \geq 0 \end{cases} \tag{3.27}$$

and the (discounted) transition probabilities by

$$p(s,a;B) = \int_{\Omega} 1(s + a - \xi \in B)\, dF(\xi) \tag{3.28}$$

In the remainder of this section we shall show that Model II is applicable under certain conditions on c, h, p, and F. This involves verification of Conditions 3.1 - 3.3 for a reference function V that is chosen appropriately. Conditions 3.1 and 3.2 are satisfied trivially since we are considering a discounted problem with costs only, so all returns are negative and hence $r^+ = 0$ (cf. (3.27) and (3.28)). Thus the only problem remaining is to determine a $\bar{V}$ such that Condition 3.3 is satisfied. This will be done first for the classical case with linear holding and shortage costs. Then a model with non-linear costs will be analyzed, under mild conditions which do not require convexity of the cost functions. Moreover, jumps in these functions are permitted as long as they are limited in magnitude. As a consequence of these more general conditions, of course, results such as the optimality of an (s,S) policy do not necessarily hold.

(a) The classical inventory model.

In this case the model as described, e.g., by Scarf [23] will be studied. The cost of ordering an amount $a \geq 0$ is given by

$$c(a) = \delta(a)\, K + c \cdot a$$

where $K \geq 0$, $c \geq 0$, and $\delta(a) = 1$, if $a > 0$, $\delta(a) = 0$, if $a = 0$. The holding cost is linear,

$$h(s) = h \cdot s, \quad s \geq 0,$$

where $h \geq 0$. The shortage cost is also linear,

$$p(-s) = -p \cdot s, \quad s < 0.$$

Hence the one-period return function can be written as

$$r(s,a) = \begin{cases} -h \cdot s - \delta(a) \cdot K - c \cdot a, & s \geq 0, \ a \geq 0 \\ p \cdot s - \delta(a) \cdot K - c \cdot a, & s < 0, \ a \geq 0 \end{cases} \tag{3.30}$$

We assume that

$$c < \rho(1-\rho)^{-1} p \tag{3.31}$$

The economic interpretation of this condition is that it costs less to order a unit now than to backlog that unit forever, which is a plausible assumption.

As an upper bound on V^* we shall use $\bar{V}$ defined by

$$\bar{V}(s) := \begin{cases} \min(-h \cdot s(1-\rho)^{-1} + h \cdot M_1 \cdot \rho(1-\rho)^{-2}, 0), & s \geq 0 \\ (p+c) \cdot s, & s < 0 \end{cases} \tag{3.32}$$

Lemma 3.7. $\quad V^* \leq \bar{V} \leq 0.$

Proof. Obviously $\bar{V} \leq 0$. We shall prove $V^* \leq \bar{V}$ inductively. Define the function b by

$$b(s) := \begin{cases} -h \cdot s, & s \geq 0 \\ p \cdot s, & s < 0 \end{cases} \tag{3.33}$$

Clearly $V^* \leq b \leq 0$, and hence

$$V^* = U^n V^* \leq U^n b \leq 0, \quad n \geq 1. \tag{3.34}$$

Define

$$h_n := h(1+\rho+\ldots+\rho^n),\ n \geq 0,$$

$$p_n := p(1+\rho+\ldots+\rho^n),\ n \geq 0, \tag{3.35}$$

$$M'_n := h \cdot M_1 \cdot \rho(1+2\rho+\ldots+n\rho^{n-1}),\ n \geq 1$$

($M'_0 := 0$). The induction hypothesis is the following.

$$U^n b(s) \leq \begin{cases} -h_n \cdot s + M'_n, & s \geq 0 \\ \max\ (p_n,\ p+c) \cdot s, & s < 0 \end{cases} \tag{3.36}$$

Noting that (3.31) implies that $p+c \leq p+\rho p_{n-1} = p_n$ for sufficiently large n, we see that the desired result will follow from (3.34) and (3.36), upon letting $n \to \infty$.

Clearly (3.36) is true for $n = 0$. Suppose that it is true for some $n \geq 0$.

Case 1 - $s \geq 0$:

$$U^{n+1} b(s) = -h \cdot s + \sup_{a \geq 0} \{-K \cdot \delta(a) - c \cdot a + \rho \int_0^\infty U^n b(s+a-\xi) d\phi(\xi)\}$$

$$\leq -h \cdot s + \sup_{a \geq 0} \{\rho \int_0^{s+a} [-h_n \cdot (s+a-\xi) + M'_n] d\phi(\xi)\}$$

$$\leq -h \cdot s + \sup_{a \geq 0} \{-\rho \cdot h_n \int_0^{s+a} (s+a-\xi) d\phi(\xi)\} + \rho M'_n$$

$$= -h \cdot s - \rho \cdot h_n \int_0^s (s-\xi) d\phi(\xi) + \rho M'_n$$

$$= -h \cdot s - \rho \cdot h_n [s\phi(s) - \int_0^s \xi d\phi(\xi)] + \rho M'_n$$

$$= -(h+\rho \cdot h_n)s + \rho \cdot h_n [s(1-\phi(s)) + \int_0^s \xi d\phi(\xi)] + \rho M'_n$$

$$\leq -h_{n+1} \cdot s + \rho \cdot h_n \cdot M_1 + \rho M'_n$$

$$= -h_{n+1} \cdot s + M'_{n+1}$$

Case 2 - $s < 0$:

Define $k = \min\{n \mid c \leq \rho p_n\}$. (Recall that (3.31) implies that $k < \infty$.) First suppose $n < k$, so that $c > \rho p_n$.

$$U^{n+1}b(s) = p \cdot s + \max\{\rho \int_0^\infty U^n b(s-\xi)d\phi(\xi),\ c \cdot s - K + \sup_{y>s}\{-c \cdot y + \rho \int_0^\infty U^n b(y-\xi)d\phi(\xi)\}\}$$

$$\leq p \cdot s + \max\{\rho \int_0^\infty p_n \cdot (s-\xi)d\phi(\xi),$$

$$c \cdot s - K + \sup_{s<y<0}\{-c \cdot y + \rho \int_0^\infty p_n \cdot (y-\xi)d\phi(\xi)\},$$

$$c \cdot s - K + \sup_{y \geq 0}\{-c \cdot y + \rho \int_y^\infty p_n \cdot (y-\xi)d\phi(\xi)\}\}$$

$$= p \cdot s + \max\{\rho \cdot p_n \cdot s - \rho \cdot p_n \cdot M_1,$$

$$c \cdot s - K + \sup_{s<y<0}\{(-c+\rho p_n) \cdot y\} - \rho \cdot p_n \cdot M_1,$$

$$c \cdot s - K - \rho \cdot p_n \cdot M_1\}$$

$$= p \cdot s - \rho \cdot p_n \cdot M_1 + \max\{\rho \cdot p_n \cdot s,\ c \cdot s - K - c \cdot s + \rho \cdot p_n \cdot s,\ c \cdot s - K\}$$

$$\leq (p + \rho \cdot p_n)s = p_{n+1} \cdot s$$

Now suppose $n = k$.

$$U^{k+1}b(s) \leq p \cdot s + \max\{\rho \int_0^\infty p_k \cdot (s-\xi)d\phi(\xi),$$

$$c \cdot s - K + \sup_{s<y<0}\{-c \cdot y + \rho \int_0^\infty p_k \cdot (y-\xi)d\phi(\xi)\},$$

$$c \cdot s - K + \sup_{y \geq 0}\{-c \cdot y + \rho \int_y^\infty p_k \cdot (y-\xi)d\phi(\xi)\}\}$$

$$\leq p \cdot s + \max\{\rho \cdot p_k \cdot s - \rho \cdot p_k \cdot M_1,$$

$$c \cdot s - K + \sup_{s<y<0}\{(-c+\rho p_k) \cdot y\} - \rho \cdot p_k \cdot M_1,$$

$$c \cdot s - K\}$$

$$\leq p\cdot s + \max\{\rho\cdot p_k\cdot s,\ c\cdot s\}$$

$$= (p+c)\cdot s$$

Finally, suppose $n > k$.

$$U^{n+1}b(s) \leq p\cdot s + \max\{\rho\int_0^\infty (p+c)\cdot(s-\xi)d\phi(\xi),$$

$$c\cdot s - K + \sup_{s<y<0}\{-c\cdot y+\rho\int_0^\infty (p+c)\cdot(y-\xi)d\phi(\xi)\},$$

$$c\cdot s - K + \sup_{y=0}\{-c\cdot y+\int_y^\infty (p+c)\cdot(y-\xi)d\phi(\xi)\}\}$$

$$\leq p\cdot s + \max\{\rho(p+c)s - \rho(p+c)M_1,$$

$$c\cdot s - K + \sup_{s<y<0}\{(-c+\rho(p+c))y\} - \rho(p+c)M_1,$$

$$c\cdot s - K\}$$

$$\leq p\cdot s + \max\{\rho(p+c)s,\ c\cdot s\}$$

$$= (p+c)\cdot s.$$

This completes the induction and so the lemma is proved.

Our main result for this model is contained in the following theorem.

Theorem 3.8. Consider the inventory-control model with linear costs and with $\bar{V}$ defined by (3.32). Then $V^* \leq \bar{V} \leq 0$, $V^*\epsilon W(\bar{V})$ and is the unique solution in $W(\bar{V})$ to the optimality equation, $v = Uv$. Moreover, for any $V_0\epsilon W(\bar{V})$, $\|V^*-U^nV_0\| \leq \rho^n\|V^*-V_0\| < \infty$, so that U^nV_0 converges to V^* uniformly and geometrically.

Proof. In order to apply Theorem 3.1, it remains to verify that Condition 3.3 holds:
$\|U\bar{V}-\bar{V}\| < \infty$. In fact, we show that $0 \leq U\bar{V}(s) - \bar{V}(s) \leq -M$, where $M < \infty$, for all $s\epsilon S$.

Case 1 - $s \geq 0$:

Using essentially the same argument as used to verify the induction hypothesis (3.36) in Lemma 3.7, it can be shown that

$$U\bar{V}(s) \leq -h \cdot s(1-\rho)^{-1} + h \cdot M_1 \cdot \rho(1-\rho)^{-2},$$

and hence $U\bar{V}(s) \leq \bar{V}(s)$. The only difference is that h_n and M_n are replaced by their respective limits, $h(1-\rho)^{-1}$ and $h \cdot M_1 \cdot \rho(1-\rho)^{-2}$. On the other hand, $U\bar{V} \geq r(f_0) + P(f_0)\bar{V}$, with f_0 the policy that has $f(s) = 0$ for $s \geq 0$, and $f(s) = -s$ for $s < 0$. Hence

$$\begin{aligned}
U\bar{V}(s) &\geq -h \cdot s + \rho \int_0^s [-h(1-\rho)^{-1}(s-\xi)]d\phi(\xi) + \rho \int_s^\infty (p+c)(s-\xi)d\phi(\xi) \\
&\geq -h \cdot s - h \cdot \rho(1-\rho)^{-1}[s\phi(s) - \int_0^s \xi d\phi(\xi)] \\
&\qquad + \rho(p+c)[s(1-\phi(s)) - \int_s^\infty \xi d\phi(\xi)] \\
&= -h \cdot s - h \cdot \rho(1-\rho)^{-1} \cdot s + [h \cdot \rho(1-\rho)^{-1} + \rho(p+c)] \cdot \\
&\qquad [s(1-\phi(s)) + \int_0^s \xi d\phi(\xi)] - \rho(p+c)M_1 \\
&\geq -h(1-\rho)^{-1}s - \rho(p+c)M_1 \\
&= \bar{V}(s) - \rho \cdot M_1(h(1-\rho)^{-2} + p + c) \\
&\geq \bar{V}(s) - M,
\end{aligned}$$

for sufficiently large $M < \infty$.

Case 2 - $s < 0$:

Again, the proof that $U\bar{V}(s) \leq \bar{V}(s) = (p+c) \cdot s$ is essentially the same as the proof of (3.36) in Lemma 3.7. On the other hand, $U\bar{V} \geq r(f_0) + P(f_0)\bar{V}$, so that

$$U\bar{V}(s) = p \cdot s - K - c(-s) + \rho \int_0^\infty (p+c)(0-\xi)d\phi(\xi)$$

$$= (p+c)\cdot s - K - \rho(p+c)M_1$$

$$\geq (p+c)\cdot s - M,$$

for sufficiently large $M < \infty$.

Remark 3.11. As in the queueing examples we can define for each decision rule f a transformed return function $\tilde{r}(f)$ by

$$\tilde{r}(f) := r(f) + P(f)\bar{V} - \bar{V} = L(f)\bar{V} - \bar{V}.$$

Now the transformed inventory control model (S, A, D, $\tilde{r}$, p) satisfied the conditions of Model I of Section 2, since $\| \sup_f \tilde{r}(f) \| = \| U\bar{V} - \bar{V} \| < M$. Moreover, $\tilde{V}^* = V^* - \bar{V}$.

Remark 3.12. The proof of Theorem 3.8 shows that $U\bar{V} \leq \bar{V}$. It follows by induction from the monotonicity of U that $U^n\bar{V} = U^{n-1}\bar{V}$, so that the convergence of $U^n\bar{V}$ to V^* is monotonically decreasing.

Remark 3.13. The idea of extremal policies as introduced for queueing systems cannot be used directly for inventory systems. It is clear from the proof of Theorem 3.8, however, that any policy f_0 of the form $f_0(s) = 0$ (do not order) for all $s > s_+ \geq 0$ and $-s-k \leq f_0(s) \leq -S$ for $s < 0$ and some $0 < k < \infty$ satisfies $\| V^* - V^{f_0} \| < \infty$, which implies that V^{f_0} might serve as a reference policy. For such f_0 we might define $\tilde{r}(f)$ by

$$\tilde{r}(f) := r(f) + P(f)V^{f_0} - V^{f_0},$$

which automatically implies that $\tilde{r}(f_0) = 0$. Consequently, the successive approximations method,

$$v_0 = 0$$

$$v_n = \sup_f \{\tilde{r}(f) + P(f)v_{n-1}\}, \quad n \geq 1,$$

converges to $\tilde{V}^*$ monotonically, i.e., $v_n \geq v_{n-1}$, and $v_n \to \tilde{V}^* = V^* - V^{f_0}$.

As in the queueing examples the effect of the shift transformation is to subtract V^{f_0} from the value function for each policy. So again the value functions are measured relative to that of a reference policy, in this case f_0. Roughly speaking, our results say that the inventory problem with linear cost structure is equivalent to one in which the costs one has to incur to reach the "feasible area" (not too far from state s = 0) are subtracted.

(b) Inventory control with restricted order quantity

We consider the same problem as described in Example 3(a) with the restriction that the maximal order quantity is R. In this model it is not always possible from states $s < 0$ to reach the "feasible area" in one step. Hence a logical reference policy will be of the form: $f_0(s) = R$ for $s < s_0 \leq 0$; $f_0(s) = 0$ for $s > s_1 \geq 0$; $f(s) \leq R$ for $s \in [s_0, s_1]$. The value function for such a policy will be of the order:

$$V(s) := \begin{cases} -h \cdot s(1-\rho)^{-1}, & \text{for } s \geq 0 \\ p \cdot s(1-\rho)^{-1}, & \text{for } s < 0. \end{cases} \tag{3.37}$$

Define $\bar{V}$ by

$$\bar{V}(s) := \begin{cases} \min(-h \cdot s(1-\rho)^{-1} + h \cdot M_1 \cdot \rho(1-\rho)^{-2}, 0), & \text{for } s \geq 0 \\ \min(p \cdot s(1-\rho)^{-1} + p \cdot R \cdot \rho(1-\rho)^{-2}, 0), & \text{for } s < 0. \end{cases} \tag{3.38}$$

Lemma 3.9. $\quad V^* \leq \bar{V} \leq 0.$

Proof. The proof parallels that of Lemma 3.7. With b again defined by (3.33), the induction hypothesis is now

$$U^n b(s) \leq \begin{cases} -h_n \cdot s + M_n', & s \geq 0 \\ p_n \cdot s + M_n'', & s < 0 \end{cases} \tag{3.39}$$

with h_n, p_n, and M_n' defined by (3.35) and

$$M_n'' := p \cdot R \cdot \rho(1+2\rho+\ldots+n\rho^{n-1}).$$

If (3.39) is true, the desired result again follows from (3.34) upon letting $n \to \infty$.

It remains to prove (3.39).

Case 1 - $s \geq 0$:

Since $\sup_{0 \leq a \leq R} \{\cdot\} \leq \sup_{a \geq 0} \{\cdot\}$, the inductive argument used in Case 1 of Lemma 3.7 can be used again here.

Case 2 - $s < 0$:

Clearly (3.39) is true for $n = 0$.

Again define $k = \min \{n | c \leq \rho p_n\}$. For $n \leq k$, the inductive argument used in Case 2 of Lemma 3.7 can be used again here to show that

$$U^n b(s) \leq p_n \cdot s \leq p_n \cdot s + M_n''$$

Suppose (3.39) is true for some $n \geq k$.

$$\begin{aligned} (s<-R) \quad U^{n+1} b(s) &\leq p \cdot s + \max\{\rho \int_0^\infty p_n \cdot (s-\xi) d\phi(\xi) + \rho \cdot M_n'', \\ &\qquad c \cdot s - K + \sup_{s < y \leq s+R} \{-c \cdot y + \rho \int_0^\infty p_n \cdot (y-\xi) d\phi(\xi)\} + \rho \cdot M_n''\} \\ &= p \cdot s + \max\{\rho \cdot p_n s, \; c \cdot s - K + \sup_{s < y \leq s+R} \{(-c+\rho \cdot p_n) y\} \\ &\qquad\qquad - \rho \cdot p_n \cdot M_1 + \rho \cdot M_n'' \\ &\leq p \cdot s + \max\{\rho \cdot p_n \cdot s, \; c \cdot s - K + (-c+\rho \cdot p_n)(s+R)\} + \rho \cdot M_n'' \end{aligned}$$

$$= p\cdot s + \max\{\rho\cdot p_n\cdot s,\ -K - cR + \rho\cdot p_n\cdot s + \rho\cdot p_n\cdot R\} + \rho\cdot M''_n$$

$$\leq (p+\rho\cdot p_n)s + \rho\cdot p_n\cdot R + \rho\cdot M''_n$$

$$= p_{n+1}\cdot s + M''_{n+1}$$

$$(-R\leq s<0)\quad U^{n+1}b(s) = p\cdot s + \max\{\rho\int_0^\infty U^n b(s-\xi)\,d\phi(\xi),\ -K + \sup_{a\geq 0}\{-c\cdot a+\rho\int_0^\infty U^n b(s+a-\xi)\,d\phi(\xi)\}\}$$

$$\leq p\cdot s + \max\{\rho\int_0^\infty p_n\cdot(s-\xi)\,d\phi(\xi) + \rho\cdot M''_n,\ 0\}$$

$$\leq p\cdot s + \max\{\rho\cdot p_n\cdot s-\rho\cdot p_n\cdot M_1,\ 0\} + \rho\cdot M''_n$$

$$\leq p\cdot s + \max\{\rho\cdot p_n\cdot s+\rho\cdot p_n\cdot R,\ 0\} + \rho\cdot M''_n$$

$$= (p+\rho\cdot p_n)s + \rho\cdot p_n\cdot R+\rho\cdot M''_n$$

$$= p_{n+1}\cdot s + M''_{n+1}.$$

This completes the inductive step and hence the proof of the lemma.

Theorem 3.10. Consider the inventory-control model with linear costs, restricted order quantity, $0\leq a\leq R<\infty$, and $\overline{V}$ defined by (3.38). Then $V^*\leq\overline{V}\leq 0$, $V^*\in W(\overline{V})$ and is the unique solution in $W(\overline{V})$ to the optimality equation, $v = Uv$. Moreover, for any $V_0\in W(\overline{V})$, $\|V^*-U^nV_0\| \leq \rho^n\|V^*-V_0\| <\infty$, so that U^nV_0 converges to V^* uniformly and geometrically.

Proof. Again we verify Condition 3.3 by showing that $0 \geq U\overline{V}(s) - \overline{V}(s) \geq -M$, where $M<\infty$, for all $s\in S$.

Case 1 - $s \geq 0$:

The proof that $U\overline{V}(s) \leq \overline{V}(s)$ is the same as in Theorem 3.8. On the other hand, $U\overline{V} \geq r(f_0) + P(f_0)\overline{V}$, with f_0 the policy that has $f(s) = 0$ for $s \geq -R$, and $f(s) = R$, for $s < -R$. Hence

$$U\bar{V}(s) \geq -h\cdot s + \rho\int_0^s [-h(1-\rho)^{-1}(s-\xi)]d\phi(\xi)$$

$$+ \rho\int_s^\infty p(1-\rho)^{-1}(s-\xi)d\phi(\xi)$$

$$\geq -h\cdot s - h\cdot\rho(1-\rho)^{-1}[s\phi(s) - \int_0^s \xi d\phi(\xi)]$$

$$+ p\cdot\rho(1-\rho)^{-1}[s(1-\phi(s)) - \int_s^\infty \xi d\phi(\xi)]$$

$$\geq -h(1-\rho)^{-1}s - p\cdot\rho(1-\rho)^{-1}M_1$$

$$\geq \bar{V}(s) - \rho(1-\rho)^{-1}M_1[h(1-\rho)^{-1} + p]$$

$$\geq \bar{V}(s) - M,$$

for sufficiently large $M < \infty$.

Case 2 - $s < 0$:

The proof that $U\bar{V}(s) \leq \bar{V}(s)$ is essentially the same as in the inductive step of the proof of Lemma 3.9 for $n \geq k$. The only difference is that p_n and M_n' are replaced by their respective limits, $p(1-\rho)^{-1}$ and $p\cdot R\cdot\rho(1-\rho)^{-2}$. On the other hand, $U\bar{V} \geq r(f_o) + P(f_o)\bar{V}$, so that

$(s < -R)$ $$U\bar{V}(s) \geq p\cdot s - K - c\cdot R + \rho\int_0^\infty p(1-\rho)^{-1}(s + R - \xi)d\phi(\xi)$$

$$= p\cdot s + p\cdot\rho(1-\rho)^{-1}s + p\cdot R\cdot\rho(1-\rho)^{-1} - K - c\cdot R - p\cdot\rho(1-\rho)^{-1}M_1$$

$$= p(1-\rho)^{-1}s + p\cdot R\cdot\rho(1-\rho)^{-2} - [K + c\cdot R + p\cdot\rho(1-\rho)^{-1}(M_1$$

$$+ R(1-\rho)^{-1})]$$

$$\geq \bar{V}(s) - M,$$

for sufficiently large $M < \infty$.

$$(-R \leq s < 0) \qquad U\bar{V}(s) \geq p \cdot s + \rho \int_0^\infty p(1-\rho)^{-1}(s-\xi)d\phi(\xi)$$

$$= p \cdot s + p \cdot \rho(1-\rho)^{-1}s - p \cdot \rho(1-\rho)^{-1}M_1$$

$$= p(1-\rho)^{-1}s + p \cdot R \cdot \rho(1-\rho)^{-2} - p \cdot R \cdot \rho(1-\rho)^{-2} - p \cdot \rho(1-\rho)^{-1}M_1$$

$$\geq \bar{V}(s) - p \cdot \rho(1-\rho)^{-1}[M_1 + R(1-\rho)^{-1}]$$

$$\geq \bar{V}(s) - M,$$

for sufficiently large $M < \infty$.

This completes the proof of Theorem 3.10.

Remark 3.14. The economic regularity condition (3.31), $c < \rho(1-\rho)^{-1}p$, is not needed in the model with restricted order quantity. In fact, if $c \geq \rho(1-\rho)^{-1}p$, then $k = \infty$ and the proof of Lemma 3.9 shows that $U^n b(s) \leq p_n \cdot s$, for all $s < 0$ and $n \geq 0$. Hence, in this case, $\bar{V}$ can be defined as $\bar{V}(s) := p(1-\rho)^{-1}s$, for $s < 0$.

Remark 3.15. It should be clear that it is not difficult in general to find a good reference function $\bar{V}$ or W. Usually the structure of the problem will indicate the direction in which to search for such a function. This was true for the queueing examples as well as for the inventory-control models.

Remark 3.16. Once again the convergence of $U^n\bar{V}$ to V^* is monotonically decreasing, since $U\bar{V} \leq \bar{V}$ for $\bar{V}$ defined by (3.38).

Remark 3.17. By defining $\tilde{r}(f) := r(f) + P(f)\bar{V} - \bar{V}$, the inventory-control problem with restricted order quantity can be transformed into a problem that satisfies the conditions for Model I.

(c) Inventory control with non-linear costs

Now we shall treat the inventory control model as described in the introduction of example 3, without the restriction to linear costs. We shall need the following condition:

$$c(-s) = \rho \cdot p(-s) + M_2, \qquad s \overset{\leq}{=} 0, \tag{3.40}$$

where $M_2 < \infty$. The economic interpretation of this condition is that below some point $s_0 \overset{\leq}{=} 0$ it cannot be much worse to order up to zero than to stay for one period with the shortage $s \overset{\leq}{=} s_0$. Note that for the case of linear costs, this condition is stronger than (3.31), but still economically plausible. In addition we assume that ordering all at once cannot be much worse than ordering separately, i.e.,

$$c(a+b) \overset{\leq}{=} c(a) + c(b) + M_3, \qquad a,b \overset{\geq}{=} 0, \tag{3.41}$$

where $M_3 < \infty$. Finally, the expected shortage cost for one stage, if we start with $s = 0$, is assumed to be finite, i.e.,

$$\int_0^\infty p(\xi)d\phi(\xi) =: M_4 < \infty \tag{3.42}$$

Define $\overline{V}$ by

$$\overline{V}(s) := \sup_{\pi} E_s^{\pi}[\sum_{t=0}^{\infty} d(X_t)] + g(s), \tag{3.43}$$

where

$$d(s) := \begin{cases} -h(s), & s \overset{\geq}{=} 0 \\ 0, & s < 0 \end{cases}$$

and

$$g(s) := \begin{cases} 0, & s \overset{\geq}{=} 0 \\ -p(-s) - c(-s), & s < 0. \end{cases}$$

Lemma 3.11. $V^* \overset{\leq}{=} \overline{V} \overset{\leq}{=} 0$.

Proof: Obviously $\overline{V} \overset{\leq}{=} 0$. On the other hand,

$(s \geq 0)$ $$\bar{V}(s) = \sup_{\pi} E_s^{\pi}[\sum_{t=0}^{\infty} d(X_t)]$$

$$\geq \sup_{\pi} E_s^{\pi}[\sum_{t=0}^{\infty} r(X_t, A_t] = V^*(s),$$

since $r(s,a) \leq d(s)$, for all $s \in S$, $a \in A$;

$(s < 0)$ $$V^*(s) = UV^*(s)$$

$$= -p(-s) + \sup_{a \geq 0} \{-c(a) + \rho \int_0^{\infty} V^*(s+a-\xi) d\phi(\xi)\}$$

$$= -p(-s) + \max \{\sup_{0 \leq a < -s} \{-c(a) + \rho \int_0^{\infty} V^*(s+a-\xi) d\phi(\xi)\},$$

$$\sup_{a \geq -s} \{-c(a) + \rho \int_0^{\infty} V^*(s+a-\xi) d\phi(\xi)\}\}$$

$$\leq -p(-s) + \max \{\sup_{0 \leq a < -s} \{-c(a) - \rho \int_0^{\infty} p(\xi-s-a) d\phi(\xi)\}, -c(-s)\}$$

$$\leq -p(-s) - c(-s) + \max \{\sup_{0 \leq a < -s} \{c(-s) - c(a) - \rho p(-s-a)\}, 0\}$$

$$\leq -p(-s) - c(-s) + \max \{\sup_{0 \leq a < -s} \{c(-s-a) - \rho p(-s-a)\} + M_3, 0\}$$

$$\leq -p(-s) - c(-s) + M_2 + M_3$$

This completes the proof of the lemma.

Theorem 3.12. Consider the inventory-control model with non-linear costs satisfying (3.40), (3.41), and (3.42), and $\bar{V}$ defined by (3.43). Then $V^* \leq \bar{V} \leq 0$, $V^* \in W(\bar{V})$ and is the unique solution in $W(\bar{V})$ to the optimality equation, $v = Uv$. Moreover, for any $V_0 \in W(\bar{V})$, $||V^* - U^n V_0|| \leq \rho^n ||V^* - V_0|| < \infty$, so that $U^n V_0$ converges to V^* uniformly and geometrically.

Proof: Let f_0 be the policy that has $f_0(s) = 0$ for $s \geq 0$, and $f_0(s) = -s$ for $s < 0$. We shall show that $||\bar{V} - V^{f_0}|| < \infty$. This will imply that $||\bar{V} - V^*|| < \infty$, $||UV^{f_0} - V^{f_0}|| < \infty$, and $||V^* - V^{f_0}|| < \infty$.

It suffices to show that $V^{f_0}(s) - \overline{V}(s) \geq -M$, for all $s \in S$, where $M < \infty$. This will be done inductively. First observe that

$$\begin{aligned}\overline{V}(s) &= \sup_{\pi} E_s^{\pi}[\sum_{t=0}^{\infty} d(X_t)] \\ &= E_s^{f_0}[\sum_{t=0}^{\infty} d(X_t)] \\ &= \lim_{n\to\infty} E_s^{f_0}[\sum_{t=0}^{n-1} d(X_t)]\end{aligned}$$

Hence $\overline{V} = \lim_{n\to\infty} \overline{V}_n$, with $\overline{V}_n$ defined by

$$\overline{V}_n(s) = \begin{cases} E_s^{f_0}[\sum_{t=0}^{n-1} d(X_t)], & s \geq 0 \\ -p(-s) - c(-s), & s < 0 \end{cases}$$

Moreover $V^{f_0} = \lim_{n\to\infty} V_n^{f_0}$, with $V_n^{f_0}$ defined by

$$V_n^{f_0}(s) := E_s^{f_0}[\sum_{t=0}^{n-1} r(X_t, A_t)]$$

The induction hypothesis is the following:

$$V_n^{f_0}(s) - \overline{V}_n(s) \geq -M, \qquad s \in S \tag{3.44}$$

Clearly (3.44) is true for n=1, since $V_1^{f_0} = \overline{V}_1$. Suppose that it is true for some $n \geq 1$.

Case 1 - $s \geq 0$:

$$\begin{aligned}V_{n+1}^{f_0}(s) &= -h(s) + \rho\int_0^{\infty} V_n^{f_0}(s-\xi)d\phi(\xi) \\ &= d(s) + \rho\int_0^{\infty} V_n^{f_0}(s-\xi)d\phi(\xi) \\ &\geq d(s) + \rho\int_0^{\infty} \overline{V}_n(s-\xi)d\phi(\xi) - \rho M \\ &\geq \overline{V}_{n+1}(s) - M.\end{aligned}$$

Case 2 - $s \geq 0$:

$$
\begin{aligned}
V^{f_0}_{n+1}(s) &= -p(-s) - c(-s) + \rho \int_0^\infty V^{f}_{n}{}_0(-\xi)\,d\phi(\xi) \\
&\geq -p(-s) - c(-s) + \rho \int_0^\infty \overline{V}_n(-\xi)\,d\phi(\xi) - \rho M \\
&= -p(-s) - c(-s) - \rho \int_0^\infty [p(\xi)+c(\xi)]\,d\phi(\xi) - \rho M \\
&\geq -p(-s) - c(-s) - \rho \int_0^\infty (1+\rho)w(\xi)\,d\phi(\xi) - \rho(M_2 + M) \\
&\geq -p(-s) - c(-s) - \rho[(1+\rho)M_4 + M_2 + M] \\
&\geq -p(-s) - c(-s) - M \\
&= \overline{V}_{n+1}(s) - M,
\end{aligned}
$$

for sufficiently large $M < \infty$.

This completes the proof of the theorem.

Remark 3.18. The inventory-control problem with general cost structure and restricted order quantity can be handled in a similar way, provided $w(-(s+R)) - p(-s) > -M$ uniformly in $s < -R$.

References

1. Bellman, R., Dynamic Programming, Princeton: Princeton University Press, 1957.

2. Bessler, S., and Veinott, A.F., Jr., "Optimal Policy for a Dynamic Multi-Echelon Inventory Model," Naval Res. Log. Quart. 13, 355-389 (1966).

3. Blackwell, D., "Discounted Dynamic Programming," Ann. Math. Statist. 36, 226-235 (1965).

4. Crabill, T.B., Gross, D., and Magazine, M.J., "A Classified Bibliography of Research on Optimal Design and Control of Queues," Opns. Res. 25, 219-232 (1977).

5. Denardo, E.V., "Contraction Mappings in the Theory Underlying Dynamic Programming," SIAM Rev. 9, 165-177 (1967).

6. Doshi, B.T., "Continuous-Time Control of the Arrival Process in an M/G/1 Queue," Stoch. Proc. and their Applications 5, 265-284 (1977).

7. Harrison, J.M., "Discrete Dynamic Programming with Unbounded Rewards," Ann. Math. Statist. 43, 636-644 (1972).

8. Hastings, N.A.J., and Mello, J., "Test for Non-Optimal Actions in Discounted Markov Programming," Management Science 19, 1019-1022 (1973).

9. Hastings, N.A.J., and van Nunen, J.A.E.E., "The Action-Elimination Algorithm for Markov Decision Processes," Markov Decision Theory (ed. H.C. Tijms & J. Wessels) Mathematical Centre Tracts 93, 161-170 (1977).

10. Hee, K.M. van, Hordijk, A., Wal, J. van der, "Successive Approximations for Convergent Dynamic Programming" Markov Decision Theory (ed. H.C. Tijms and J. Wessels), Mathematical Centre Tract 93, 183-212 (1977).

11. Hinderer, K., Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter, Lecture Notes in Operations Research and Mathematical Systems 33. New York, Springer-Verlag, 1970.

12. Johansen, S.G., and Stidham, S.,Jr., "Control of Arrivals to a Stochastic Input-Output System," NCSU-IE Tech. Rpt. No. 78-1, Dept. of Industrial Engineering, North Carolina State University, April, 1978 (forthcoming in Applied Probability).

13. Lehman, E., "Ordered Families of Distributions," Ann. Math. Statist. 25, 339-419 (1955).

14. Lippman, S.A., "On Dynamic Programming with Unbounded Rewards," Management Sci. 21, 1225-1233 (1975).

15. Lippman, S.A., "Applying a New Device in the Optimization of Exponential Queueing Systems," Opns. Res. 23, 687-710 (1975).

16. Lippman, S.A., and Stidham, S.,Jr., "Individual versus Social Optimization in Exponential Congestion Systems," Opns Res. 25, 233-247 (1977).

17. MacQueen, J., A Modified Dynamic Programming Method for Markovian Decision Problems," J. Math. Anal. Appl. 14, 38-43 (1966).

18. MacQueen, J., "A Test for Suboptimal Actions in Markovian Decision Problems," Opns. Res. 15, 559-561 (1967).

19 Morton, T.E., and Wecker, W.E., Discounting, Ergodicity, and Convergence for Markov Decision Processes," Management Sci. 23, 890-900 (1977).

20. Nunen, J.A.E.E. van, Contracting Markov Decision Processes, Mathematical Centre Tracts 71, (1976).

21. Nunen, J.A.E.E. van, Wessels, J., "Markov Decision Processes with Unbounded Rewards," Markov Decision Theory (ed. H.D. Tijms and J. Wessels), Mathematical Centre Tract 93, 1-24, (1977).

22. Porteus, E.L., "Bounds and Transformations for Discounted Finite Markov Decision Chains," Opns. Res. 23, 761-784 (1975).

23. Scarf, H., "The Optimality of (S,s) Policies in the Dynamic Inventory Problem, Chap. 13 in Mathematical Methods in the Social Sciences (ed. K.J. Arrow, S. Karlin, and P. Suppes) Stanford: Stanford University Press, 1960.

24. Schäl, M., "Conditions for Optimality in Dynamic Programming and for the Limit of n-Stage Optimal Policies to be Optimal," Z. Wahrscheinlichkeitstheorie verw. Geb. 32, 179-196 (1975).

25. Schleef, H., "Optimal Control Models for Multi-Server Exponential Queueing Systems," unpublished Ph.D. Dissertation, University of Chicago (1977)

26. Serfozo, R.F., "Monotone Optimal Policies for Markov Decision Processes," Mathematical Programming Study 6, 202-215, North-Holland (1976).

27. Shreve, S.E., and Bertsekas, D.P., "Universally Measurable Policies in Dynamic Programming," Math. of Opns. Res. 4, 15-30, (1979)

28. Sobel, M.F., "Optimal Operation of Queues," Mathematical Methods in Queueing Theory, Lecture Notes in Economics and Mathematical Systems, 98, 263-294 New York: Springer-Verlag, 1974.

29. Sobel, M.F., "The Optimality of Full-Service Policies," paper presented at Symposium on Stochastic Systems, University of Kentucky, Lexington, Kentucky, June, 1975.

30. Sobel, M.F., and Winston, W., "Optimal Extremal Congestion Management Policies," School of Organization and Management, Yale University, 1977 (revised).

31. Stidham, S , Jr., Socially and Individually Optimal Control of Arrivals to a GI/M/1 Queue," Management Sci. 24, 1598-1610 (1978).

32. Stidham, S., Jr., "On the Convergence of Successive Approximations in Dynamic Programming with Non-Zero Terminal Reward," NCSU-IE Technical Report No. 78-9, Department of Industrial Engineering, North Carolina State University, October 1978 (forthcoming in Z. für Operations Research .

33. Stidham, S., Jr., and Prabhu, N.U., "Optimal Control of Queues", Mathematical Methods in Queueing Theory, Lecture Notes in Economics and Mathematical Systems, 98, 263-293, New York: Springer-Verlag, 1974.

34. Strauch, R.E., "Negative Dynamic Programming," Ann. of Math. Statist. 37, 871-890 (1966).

35. Topkis, D., "Minimizing a Submodular Function on a Lattice," Opns. Res. 26, 305-321 (1978)

36. Veinott, A. F., Jr., Unpublished class notes, Stanford University, 1967.

37. Wessels, J., "Markov Programming by Successive Approximations with respect to Weighted Supremum Norms," J. Math. Anal. Appl. 58 (1977).

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1 REPORT NUMBER: #96	2 GOVT ACCESSION NO.	3 RECIPIENT'S CATALOG NUMBER
4 TITLE (and Subtitle): THE SHIFT-FUNCTION APPROACH FOR MARKOV DECISION PROCESSES WITH UNBOUNDED RETURNS	5 TYPE OF REPORT & PERIOD COVERED: TECHNICAL REPORT
	6 PERFORMING ORG. REPORT NUMBER
7 AUTHOR(s): Shaler Stidham, Jr. and Jo van Nunen	8 CONTRACT OR GRANT NUMBER(s): N00014-76-C-0418
9 PERFORMING ORGANIZATION NAME AND ADDRESS: OPERATIONS RESEARCH PROGRAM -ONR DEPARTMENT OF OPERATIONS RESEARCH STANFORD UNIVERSITY, STANFORD, CALIF.	10 PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (NR-047-061)
11 CONTROLLING OFFICE NAME AND ADDRESS: OFFICE OF NAVAL RESEARCH OPERATIONS RESEARCH PROGRAM CODE 434 ARLINGTON, VA. 22217	12 REPORT DATE: JULY 1981 13 NUMBER OF PAGES: 53
14 MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15 SECURITY CLASS. (of this report): UNCLASSIFIED 15a DECLASSIFICATION DOWNGRADING SCHEDULE

16 DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17 DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18 SUPPLEMENTARY NOTES

Also issued as Technical Report No. 60 Dept. of Operations Research, Stanford University under NSF GRANT ECS80-17867.

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

Dynamic programming, Markov decision processes,
Queueing theory, Control of queues, Inventory control

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

PLEASE SEE OTHER SIDE

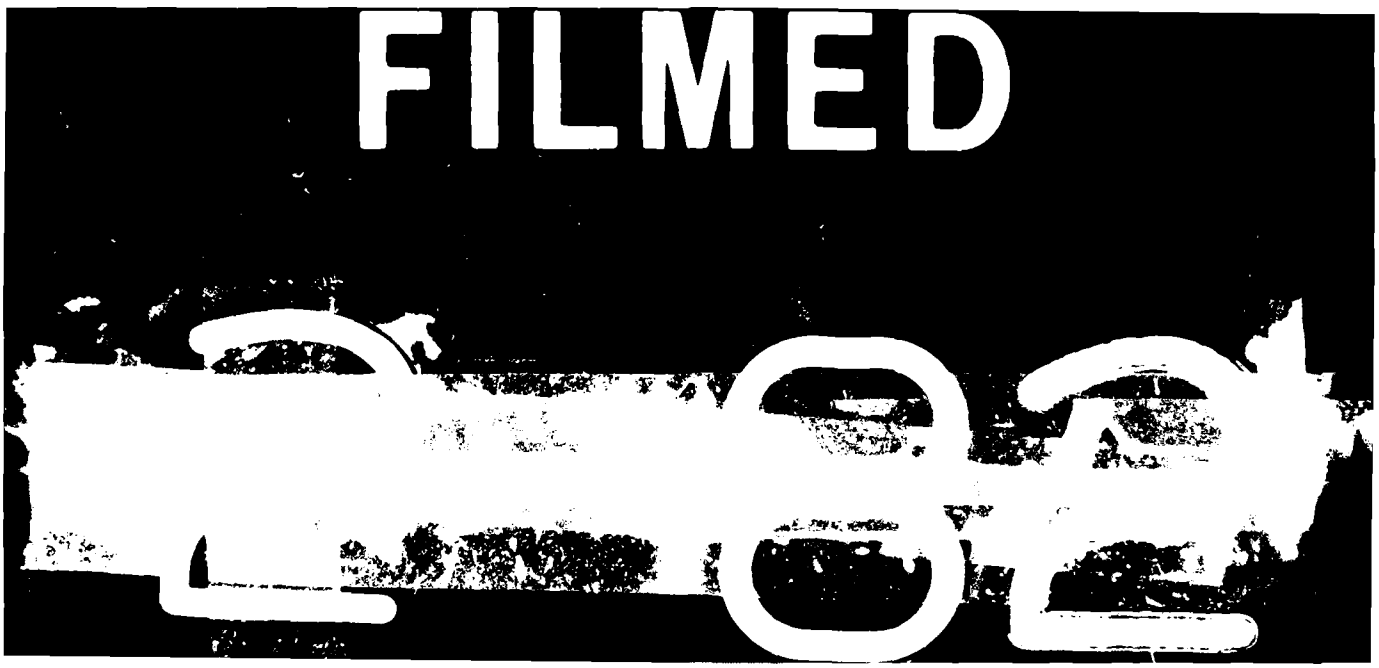
DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 68 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

TECHNICAL REPORT NO. 98 - Authors: Shaler Stidham, Jr. and Jo van Nunen

We study a discrete-time Markov decision process with general state and action space. The objective is to maximize the expected total return over a finite or infinite horizon. The transition probability measure is allowed to be defective, so that the model includes discounting, state-and action-dependent transition times (semi-Markov decision processes), and stopping problems. With applications to control of queues and inventory systems as a motivation, we develop a set of conditions on the one-period return function, the transition probabilities and the terminal value function that guarantee uniform convergence (with respect to the sup norm) of the finite-horizon optimal value functions to the infinite-horizon optimal value function (successive approximations). These conditions are substantially weaker and more realistic for the applications we have in mind than those of the classical, discounted bounded model.


FILMED